Глава 9

Вероятностные модели

Третье и последнее семейство моделей машинного обучения, рассматриваемое в этой книге, – вероятностные модели. Мы уже видели, как полезны могут быть вероятности для выражения ожиданий модели относительно класса предъявленного объекта. Например, в ☞ *дереве оценивания вероятностей* (раздел 5.2) к каждому листу присоединено распределение вероятностей классов, а объект, попавший в определенный лист древовидной модели, помечается хранящимся в нем распределением. Аналогично калиброванная линейная модель преобразует расстояние до решающей границы в вероятность класса (раздел 7.4). Это все примеры так называемых *дискриминантных* вероятностных моделей. Они моделируют апостериорное распределение вероятностей $P(Y|X)$, где Y – целевая переменная, а X – признаки. Иначе говоря, они возвращают распределение вероятностей Y при заданном X.

Другой важный класс вероятностных моделей – *порождающие* модели. Они моделируют совместное распределение $P(Y, X)$ целевой переменной Y и вектора признаков X. Зная совместное распределение, мы можем вывести любое условное или маргинальное распределение с участием тех же случайных величин. В частности, поскольку $P(X) = \Sigma_y P(Y = y, X)$, то апостериорное распределение можно получить как

$$P(Y \mid X) = \frac{P(Y,X)}{\Sigma_y P(Y = y,X)}.$$

С другой стороны, порождающие модели можно описать функцией правдоподобия $P(X|Y)$, так как $P(Y, X) = P(X|Y)P(Y)$, а распределение целевой переменной, или априорное распределение, можно легко оценить или постулировать. Такие модели называются «порождающими», потому что мы можем сделать выборку из совместного распределения для получения новых данных вместе с их метками. Альтернативно можно использовать $P(Y)$ для выборки класса и $P(X|Y)$ для выборки объекта этого класса – это было показано на примере почтового спама (стр. 41). Напротив, дискриминантная модель, например дерево оценивания вероятностей или линейный классификатор, моделирует $P(Y|X)$, но не $P(X)$, и потому может использоваться для пометки данных, но не для порождения новых.

Поскольку порождающие модели умеют делать все то же, что дискриминантные, они могут показаться более предпочтительными. Однако же они не лише-

ны недостатков. Прежде всего отметим, что объем памяти, необходимой для хранения совместного распределения, экспоненциально растет вместе с числом признаков. Поэтому приходится делать упрощающие предположения, например о независимости признаков. А если такие предположения в конкретной предметной области неверны, то результаты могут оказаться неточными. Самое распространенное критическое замечание, выдвигаемое против порождающих моделей, – это то, что точность моделирования $P(X)$ на самом деле, возможно, достигается ценой менее точного моделирования $P(Y|X)$. Однако в этом вопросе еще нет окончательной ясности, и, безусловно, существуют ситуации, когда знание $P(X)$ дает желанное дополнительное понимание предметной области. Например, нас, скорее всего, не так уж и заботит неправильная классификация тех объектов, которые маловероятны согласно распределению $P(X)$.

Одна из самых привлекательных черт вероятностного взгляда на вещи – тот факт, что он позволяет рассматривать обучение как процесс уменьшения неопределенности. Например, если априорное распределение по классам равномерно, то до получения каких-либо знаний о классифицируемом объекте неопределенность в части того, какому классу он принадлежит, максимальна. Если же апостериорное распределение после наблюдения объекта не столь равномерно, значит, мы уменьшили эту неопределенность. Этот процесс можно повторять всякий раз после прихода новой информации, используя апостериорное распределение, полученное на предыдущем шаге, как априорное для следующего. В принципе, этот процесс применим к любой неизвестной величине.

Пример 9.1 (спам или неспам?). Допустим, мы хотим оценить вероятность θ того, что произвольно взятое почтовое сообщение – спам, чтобы можно было использовать подходящее априорное распределение. Напрашивающееся решение – проверить n сообщений, определить, сколько из них спамных – d, и положить $\hat{\theta} = d/n$; для этого не нужны никакие сложные статистики. Однако, хотя это наиболее правдоподобная оценка θ – оценка апостериорного максимума (MAP), пользуясь терминологией, введенной на стр. 40, это не означает, что все прочие значения θ следует полностью исключить. Мы моделируем это распределением вероятности θ, которое обновляется всякий раз при поступлении новой информации. На рис. 9.1 показано распределение, которое все сильнее и сильнее смещается в сторону спама.

Явное моделирование апостериорного распределения параметра θ обладает рядом достоинств, которые обычно ассоциируются с «байесовским» взглядом.

- Мы можем точно охарактеризовать оставшуюся неопределенность оценки, количественно оценив размах апостериорного распределения.
- Мы можем получить порождающую модель для параметра, произведя выборку из апостериорного распределения, которая содержит намного больше информации, чем может предложить сводная статистика типа MAP, – так, вместо использования единственного сообщения с $\theta = \theta_{MAP}$ порождающая модель может содержать ряд сообщений с θ, полученными в результате выборки из апостериорного распределения.

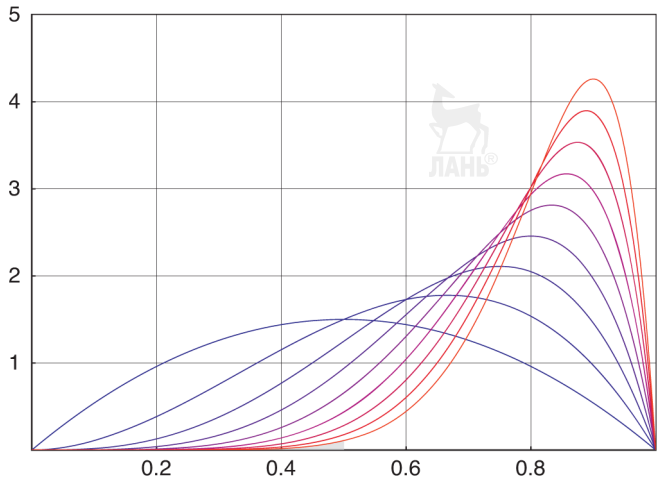

Рис. 9.1. При проверке каждого почтового сообщения мы уменьшаем неопределенность, связанную с априорной вероятностью спама θ. После просмотра двух сообщений, из которых одно оказалось спамом, возможные значения θ характеризуются симметричным распределением с центром 1/2. Если мы проверим три, четыре, ..., десять сообщений, и все они окажутся спамом, то распределение сужается и сдвигается все правее. Было бы естественно ожидать, что распределение для n сообщений достигает максимума при $\hat{\theta}_{\text{MAP}} = (n-1)/n$ (например, $\hat{\theta}_{\text{MAP}} = 0.8$ при $n = 5$); однако асимметричные распределения, подобные этому, содержат информацию, которую невозможно выразить одним числом, скажем, средним или максимумом

- Мы можем количественно оценить вероятность утверждений типа «сообщения смещены в сторону хороших» (крохотная заштрихованная область на рис. 9.1 показывает, что после наблюдения одного хорошего и девяти спамных сообщений эта вероятность очень мала – примерно 0.6%).
- Мы можем использовать одно из этих распределений для описания априорных гипотез: например, если мы полагаем, что доли спама и неспама 50–50, то можем взять в качестве априорного распределение для $n = 2$ (на рис. 9.1 оно самое нижнее, симметричное)[1].

Ключевой момент – что *вероятности необязательно интерпретировать как оценки относительных частот, они могут нести и более общий смысл: степень доверия (возможно, субъективная)*. Следовательно, мы можем связать распределение вероятностей практически с чем угодно: не только с признаками и целями,

[1] Статистики называют априорное распределение, имеющее такую же форму, как апостериорное, *сопряженным априорным* – в данном случае мы использовали бета-распределение, сопряженное биномиальному. Сопряженные априорные распределения не только упрощают математический аппарат, но также допускают интуитивно более понятные интерпретации; в данном случае мы делаем вид, что уже проверили два сообщения, одно из которых оказалось спамом, – очень полезная идея, которую мы уже фактически применяли в разделе 2.3 в форме поправки Лапласа.

но и с параметрами моделей и даже самими моделями. Например, в только что приведенном примере мы рассматривали распределение $P(\theta|D)$, где D представляет данные (то есть классы проверяемых почтовых сообщений).

С вероятностными моделями связано важное понятие *оптимальности по Байесу*. Классификатор называется оптимальным по Байесу, если он всегда назначает объекту x класс $\text{argmax}_y P^*(Y = y|X = x)$, где P^* обозначает истинное апостериорное распределение. И хотя на практике истинное распределение почти никогда не известно, существует несколько способов конкретизации этого понятия. Например, мы можем поставить эксперименты с искусственно сгенерированными данными, для которых сами выберем истинное распределение: это позволяет экспериментально оценить, насколько качество модели близко к оптимальному по Байесу. С другой стороны, при выводе вероятностного метода обучения обычно делаются предположения об истинном распределении, которые позволяют теоретически доказать, что модель будет оптимальной по Байесу, если эти предположения выполняются. Например, ниже в этой главе мы сформулируем условия, при которых базовый линейный классификатор является оптимальным по Байесу. Таким образом, это свойство лучше всего рассматривать как мерило качества вероятностных моделей.

Поскольку многие модели, рассмотренные в предыдущих главах, способны оценивать вероятности классов и потому являются дискриминантными вероятностными моделями, стоит отметить, что выбор конкретной модели, который часто так и называют – *выбор модели*, необязательно означает байесовскую оптимальность – даже если выбранная модель оказывается наилучшей для истинного распределения. Для иллюстрации предположим, что m^* – наилучшее дерево оценивания вероятностей, которое мы обучили на достаточном объеме данных. С помощью m^* мы могли бы предсказать класс $\text{argmax}_y P(Y = y|M = m^*, X = x)$ для объекта x, где M – случайная величина над классом моделей, из которого выбрана модель m^*. Однако эти предсказания необязательно оптимальны по Байесу, потому что

$$\begin{aligned} P(Y|X = x) &= \textstyle\sum_{m\in M} P(Y, M = m|X = x) && \text{путем маргинализации по } M; \\ &= \textstyle\sum_{m\in M} P(Y|M = m, X = x)P(M = m|X = x) && \text{по цепному правилу;} \\ &= \textstyle\sum_{m\in M} P(Y|M = m, X = x)P(M = m) && \text{в силу независимости } M \text{ и } X. \end{aligned}$$

Здесь $P(M)$ можно интерпретировать как апостериорное распределение моделей после наблюдения обучающих данных (следовательно, модель MAP имеет вид $m^* = \text{argmax}_m P(M = m)$). Последнее выражение в приведенном выше выводе означает, что мы должны усреднить предсказания по всем моделям, приписав им веса в соответствии с апостериорными вероятностями. Очевидно, что это распределение совпадает с $P(Y|M = m^*, X = x)$, только если $P(M)$ равно нулю для всех моделей, кроме m^*, то есть если мы видели достаточно обучающих данных, чтобы исключить все модели, кроме одной. Понятно, что это нереалистичное предположение[1].

[1] Отметим, что на самом деле не требуется, чтобы оба распределения совпадали, достаточно, чтобы у них был одинаковый максимум.

Эта глава устроена следующим образом. В разделе 9.1 мы увидим некоторые полезные связи между геометрическим и вероятностным взглядом на вещи, которые проявляются, когда признаки имеют нормальное распределение. Как уже отмечалось, это позволит нам сформулировать условия, при которых базовый линейный классификатор является оптимальным по Байесу. В разделе 8.2 мы рассмотрим случай категориальных признаков, который ведет к хорошо известному наивному байесовскому классификатору. В разделе 9.3 мы вернемся к линейному классификатору, но взглянем на него с вероятностной точки зрения, это позволит разработать новый алгоритм обучения с явно поставленной целью оптимизировать апостериорную вероятность примеров. В разделе 9.4 обсуждаются способы учета скрытых переменных. Наконец, в разделе 9.5 мы вкратце познакомимся с методами обучения на основе сжатия, которым можно придать вероятностную интерпретацию с помощью понятий из теории информации.

9.1 Нормальное распределение и его геометрические интерпретации

Мы можем установить связь между вероятностными и геометрическими моделями, рассмотрев распределения вероятности, определенные в евклидовых пространствах. Самыми известными из них являются *нормальные распределения*, называемые также *гауссианами* (в замечании 9.1 приведены важнейшие факты об одномерных и многомерных нормальных распределениях). Начнем с рассмотрения одномерного случая с двумя классами. Предположим, что значения $x \in \mathbb{R}$ следуют *смесовой модели* (mixture model), то есть у каждого класса имеется собственное распределение вероятности (*компонент* смесовой модели). Будем предполагать гауссову смесовую модель, когда обе компоненты смеси – гауссианы. Тогда имеем:

$$P(x|\oplus) = \frac{1}{\sqrt{2\pi}\sigma^{\oplus}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu^{\oplus}}{\sigma^{\oplus}}\right]^2\right);$$

$$P(x|\ominus) = \frac{1}{\sqrt{2\pi}\sigma^{\ominus}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu^{\ominus}}{\sigma^{\ominus}}\right]^2\right),$$

где $\mu^{\oplus}$ и $\Sigma^{\oplus}$ – среднее и стандартное отклонения положительного класса, а $\mu^{\ominus}$ и $\Sigma^{\ominus}$ – среднее и стандартное отклонения отрицательного класса. Это дает следующее отношение правдоподобия:

$$\mathrm{LR}(x) = \frac{P(x|\oplus)}{P(x|\ominus)} = \frac{\sigma^{\ominus}}{\sigma^{\oplus}} \exp\left(-\frac{1}{2}\left[\left(\frac{x-\mu^{\oplus}}{\sigma^{\oplus}}\right)^2 - \left(\frac{x-\mu^{\ominus}}{\sigma^{\ominus}}\right)^2\right]\right). \tag{9.1}$$

Одномерное нормальное, или гауссово, распределение имеет следующую функцию плотности вероятности:

$$P(x\,|\,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)=\frac{1}{E}\exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2\right)=\frac{1}{E}\exp(-z^2/2),\ \ E=\sqrt{2\pi}\sigma.$$

Это распределение имеет два параметра: μ – среднее, или математическое, ожидание, оно же медиана (то есть точка, которая делит пополам площадь под кривой функции плотности) и мода (то есть точка, в которой функция плотности достигает максимума), и σ – стандартное отклонение, определяющее ширину колоколообразной кривой.

$z = (x - \mu)/\sigma$ называется *z-оценкой*, ассоциированной с x; она измеряет количество стандартных отклонений, укладывающихся между x и средним (сама она имеет среднее 0 и стандартное отклонение 1). Отсюда следует, что $P(x|\mu,\sigma) = (1/\sigma)P(z|0,1)$, где $P(z|0,1)$ обозначает *стандартное нормальное распределение*. Иными словами, любое нормальное распределение можно получить из стандартного нормального распределения, выполнив масштабирование по оси x с коэффициентом σ, масштабирование по оси y с коэффициентом $1/\sigma$ (чтобы площадь под кривой оставалась равной 1) и параллельный перенос начала координат на μ.

Многомерное нормальное распределение d-мерных векторов $\mathbf{x} = (x_1, ..., x_d)^T \in \mathbb{R}^d$ имеет вид:

$$P(\mathbf{x}\,|\,\mu,\boldsymbol{\Sigma})=\frac{1}{E_d}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right),\ \ E_d=(2\pi)^{d/2}\sqrt{|\boldsymbol{\Sigma}|}. \tag{9.2}$$

Его параметрами являются средний вектор $\mu = (\mu_1, ..., \mu_d)^T$ и ковариационная матрица $\boldsymbol{\Sigma}$ размерности $d \times d$ (см. замечание 7.2 на стр. 211). $\boldsymbol{\Sigma}^{-1}$ – обратная ковариационная матрица, $|\boldsymbol{\Sigma}|$ – ее определитель. Можно считать, что компонентами вектора $\mathbf{x}$ являются d признаков, возможно, коррелированных.

Если $d = 1$, то $\boldsymbol{\Sigma} = \sigma^2 = |\boldsymbol{\Sigma}|$ и $\boldsymbol{\Sigma}^{-1} = 1/\sigma^2$, что дает одномерную гауссиану в качестве частного случая.

Для $d = 2$ имеем $\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}$, $|\boldsymbol{\Sigma}| = \sigma_1^2\sigma_2^2 - (\sigma_{12})^2$ и $\boldsymbol{\Sigma}^{-1}=\frac{1}{|\boldsymbol{\Sigma}|}\begin{pmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{pmatrix}$. Применяя z-оценки, мы можем вывести следующее выражение для двумерного нормального распределения:

$$P(x_1,x_2\,|\,\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)=\frac{1}{E_2}\exp\left(-\frac{1}{2(1-\rho^2)}(z_1^2+z_2^2-2\rho z_1 z_2)\right),\ \ E_2=2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}, \tag{9.3}$$

где $z_i = (x_i - \mu_i)/\sigma_i$ для $i = 1, 2$, а $\rho = \sigma_{12}/\sigma_1\sigma_2$ – *коэффициент корреляции* между двумя признаками. Для *многомерного стандартного нормального распределения* $\mu = \mathbf{0}$ (d-мерный вектор, состоящий из одних нулей) и $\boldsymbol{\Sigma} = \mathbf{I}$ (единичная матрица размерности $d \times d$), и, следовательно, $P(\mathbf{x}|\mathbf{0},\mathbf{I}) = \frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}\mathbf{x}\cdot\mathbf{x}\right)$.

Замечание 9.1. Нормальное распределение

Рассмотрим сначала случай, когда обе компоненты имеют одинаковое стандартное отклонение, то есть $\sigma^{\oplus} = \sigma^{\ominus} = \sigma$. Тогда аргумент экспоненциальной функции в уравнении (9.1) можно упростить следующим образом:

$$-\frac{1}{2\sigma^2}[(x-\mu^{\oplus})^2-(x-\mu^{\ominus})^2]=-\frac{1}{2\sigma^2}[x^2-2\mu^{\oplus}x+\mu^{\oplus^2}-(x^2-2\mu^{\ominus}x+\mu^{\ominus^2})]=$$
$$=-\frac{1}{2\sigma^2}[-2(\mu^{\oplus}-\mu^{\ominus})x+(\mu^{\oplus^2}-\mu^{\ominus^2})]=\frac{\mu^{\oplus}-\mu^{\ominus}}{\sigma^2}\left[x-\frac{\mu^{\oplus}+\mu^{\ominus}}{2}\right].$$

Следовательно, отношение максимального правдоподобия можно записать в виде $\mathrm{LR}(x) = \exp(\gamma(x - \mu))$ с параметрами $\gamma = (\mu^{\oplus} - \mu^{\ominus})/\sigma^2$ – разность между средними, поделенная на дисперсию, и $\mu = (\mu^{\oplus} + \mu^{\ominus})/2$ – среднее арифметическое двух средних классов. Отсюда следует, что порог принятия решения с максимальным правдоподобием (значение x такое, что $\mathrm{LR}(x) = 1$) равен $x_{\mathrm{ML}} = \mu$.

Если $\sigma^{\oplus} \neq \sigma^{\ominus}$, то члены, содержащие x^2, в уравнении (9.1) не сокращаются. В результате мы получаем две решающие границы и разрывную решающую область для одного из классов.

Пример 9.2 (одномерная смесовая модель с неравными дисперсиями). Предположим, что $\mu^{\oplus} = 1$, $\mu^{\ominus} = 2$ и $\sigma^{\ominus} = 2\sigma^{\oplus} = 2$. Тогда $\mathrm{LR}(x) = 2\exp(-[(x-1)^2 - (x-2)^2/4]/2) = 2\exp(3x^2/8)$. Отсюда следует, что решающие границы с максимальным правдоподобием есть $x = \pm(8/3)\ln 2 = \pm 1.85$. На рис. 9.2 видно, что это точки, в которых пересекаются две гауссианы. Если же $\sigma^{\ominus} = \sigma^{\oplus}$, то получается единственная решающая граница с максимальным правдоподобием при $x = 1.5$.

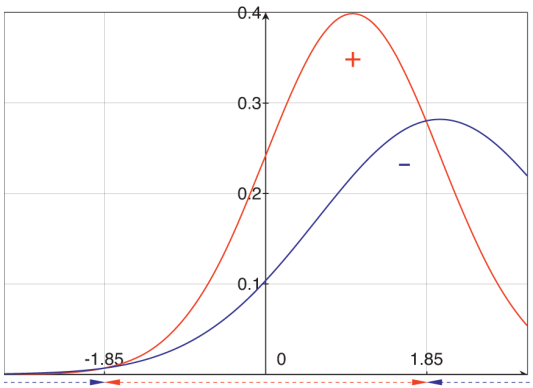

Рис. 9.2. Если положительные примеры лежат на гауссиане со средним и стандартным отклонением 1, а отрицательные – на гауссиане со средним и стандартным отклонением 2, то оба распределения пересекаются в точках $x = \pm 1.85$. Это означает, что областью максимального правдоподобия для положительных примеров является замкнутый отрезок $[-1.85, 1.85]$, а значит, соответствующая область для отрицательных примеров разрывна

Разрывные решающие области встречаются также в многомерных пространствах. В следующем примере это продемонстрировано для $m = 2$.

Пример 9.3 (двумерная гауссова смесь). Мы воспользуемся уравнением (9.3) для получения явных выражений для решающей границы с максимальным правдоподобием в двумерном случае. В этом примере мы будем считать, что $\mu_1^{\oplus} = \mu_2^{\oplus} = 1$ и $\mu_1^{\ominus} = \mu_2^{\ominus} = -1$.

(*i*) Если все дисперсии равны 1 и обе корреляции равны 0, то решающая граница с максимальным правдоподобием описывается уравнением $(x_1 - 1)^2 + (x_2 - 1)^2 - (x_1 + 1)^2 - (x_2 + 1)^2 = -2x_1 - 2x_2 - 2x_1 - 2x_2 = 0$, то есть $x_1 + x_2 = 0$ (рис. 9.3 слева).

(*ii*) Если $\sigma_1^{\oplus} = \sigma_1^{\ominus} = 1$, $\sigma_2^{\oplus} = \sigma_2^{\ominus} = \sqrt{2}$ и $\rho^{\oplus} = \rho^{\ominus} = \sqrt{2}/2$, то решающая граница с максимальным правдоподобием описывается уравнением $(x_1 - 1)^2 + (x_2 - 1)^2/2 - \sqrt{2}(x_1 - 1)(x_2 - 1)/\sqrt{2} - (x_1 + 1)^2 - (x_2 + 1)^2/2 + \sqrt{2}(x_1 + 1)(x_2 + 1)/\sqrt{2} = -2x_1 = 0$ (рис. 9.3 в центре).
(*iii*) Если все дисперсии равны 1 и $\rho^{\oplus} = \rho^{\ominus} = \rho$, то решающая граница с максимальным правдоподобием описывается уравнением $(x_1 - 1)^2 + (x_2 - 1)^2 - 2\rho(x_1 - 1)(x_2 - 1) - (x_1 + 1)^2 - (x_2 + 1)^2 - 2\rho(x_1 + 1)(x_2 + 1) = -4x_1 - 4x_2 - 4\rho x_1 x_2 - 4\rho = 0$, то есть $x_1 + x_2 + \rho x_1 x_2 + \rho = 0$, и является гиперболой. На рис. 9.3 справа это показано для $\rho = 0.7$. Отметим, что левая нижняя часть пространства объектов – положительная решающая область, хотя она не содержит ни одного обучающего примера и ближе к среднему отрицательного, чем к среднему положительного класса.

Обратите внимание на окружности и эллипсы на рис. 9.3, которые дают наглядное представление о ковариационной матрице. Проецируя фигуру, описывающую положительный класс, на ось x, мы получаем отрезок $[\mu_1^{\oplus} - \sigma_1^{\oplus}, \mu_1^{\oplus} + \sigma_1^{\oplus}]$ шириной в одно стандартное отклонение вокруг среднего – и аналогично для отрицательного класса и оси y. Можно выделить три случая: (*i*) стандартные отклонения x и y равны, и коэффициент корреляции равен нулю, тогда получается окружность; (*ii*) стандартные отклонения различны, и коэффициент корреляции равен нулю, тогда получается эллипс, параллельный оси с наибольшим стандартным отклонением; (*iii*) коэффициент корреляции не равен нулю: ориентация эллипса дает знак коэффициента корреляции, а его ширина зависит от абсолютной величины коэффициента корреляции[1]. Математически эти фигуры определяются приравниванием $f(\mathbf{x})$ в выражении $(1/E_d)\exp(-{}^1/_2 f(\mathbf{x}))$ к единице и решением относительно $\mathbf{x}$, чтобы найти точки, находящиеся на расстоянии одного стандартного отклонения от среднего. В двумерном случае это приводит к уравнению $(z_1^2 + z_2^2 - 2\rho z_1 z_2) = 1 - \rho^2$, которое можно преобразовать в уравнение эллипса относительно x_1 и x_2, если раскрыть z-оценки. Отметим, что при $\rho = 0$ получается окружность с центром в начале координат, а когда $\rho \to 1$, эта окружность стремится к прямой $z_2 = z_1$ (мы не можем положить $\rho = 1$, потому что это приведет к сингулярной ковариационной матрице).

В общем многомерном случае условие $(\mathbf{x} - \mu)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu) = 1$ определяет гиперэллипс, поскольку матрица $\boldsymbol{\Sigma}^{-1}$ обладает определенными свойствами[2]. Для стандартного нормального распределения линии с единичным стандартным отклонением лежат на гиперсфере (d-мерном аналоге окружности), определенной уравнением $\mathbf{x} \cdot \mathbf{x} = 1$. Очень полезное геометрическое соображение заключается в том, что точно так же, как гиперсферу можно преобразовать в произвольный гиперэллипс масштабированием и поворотом, любую многомерную гауссиану можно получить из стандартной гауссианы масштабированием, поворотом (для

[1] Типичная ошибка – думать, что угол поворота эллипса зависит от коэффициента корреляции; на самом деле он определяется исключительно отношением абсолютных величин маргинальных стандартных отклонений.

[2] Точнее, $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ определяет гиперэллипс, если матрица $\mathbf{A}$ симметричная и положительно определенная. Оба эти свойства удовлетворены, если $\mathbf{A}$ – матрица, обратная к несингулярной ковариационной матрице.

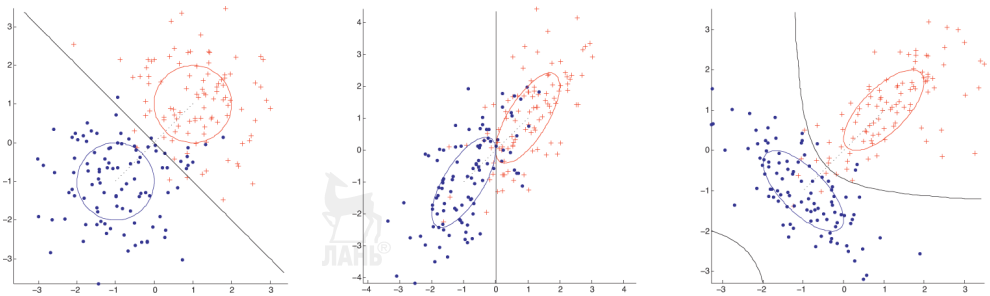

Рис. 9.3. (Слева) Если признаки некоррелированы и имеют одинаковую дисперсию, то классификация по максимальному правдоподобию приводит к базовому линейному классификатору, решающая граница которого ортогональна прямой, соединяющей средние. **(В центре)** При условии, что ковариационные матрицы идентичны, оптимальная по Байесу решающая граница линейна – если бы мы устранили корреляцию с помощью поворота и масштабирования, то снова получили бы базовый линейный классификатор. **(Справа)** При различных ковариационных матрицах получаются гиперболические решающие границы, то есть одна из решающих областей разрывна

получения требуемой ковариационной матрицы) и параллельным переносом (для получения желаемого среднего). Обратно произвольную многомерную гауссиану можно преобразовать в стандартное нормальное распределение с помощью параллельного переноса, поворота и масштабирования, как уже было сказано в замечании 1.2 на стр. 36. Это приводит к устранению корреляции признаков и их нормировке.

Общую формулу для отношения правдоподобия можно вывести из уравнения (9.2) в виде:

$$\mathrm{LR}(\mathbf{x}) = \sqrt{\frac{|\boldsymbol{\Sigma}^{\ominus}|}{|\boldsymbol{\Sigma}^{\oplus}|}} \exp\left(-\frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu}^{\oplus})^{\mathrm{T}}(\boldsymbol{\Sigma}^{\oplus})^{-1}(\mathbf{x}-\boldsymbol{\mu}^{\oplus}) - (\mathbf{x}-\boldsymbol{\mu}^{\ominus})^{\mathrm{T}}(\boldsymbol{\Sigma}^{\ominus})^{-1}(\mathbf{x}-\boldsymbol{\mu}^{\ominus})] \right),$$

где $\boldsymbol{\mu}^{\oplus}$ и $\boldsymbol{\mu}^{\ominus}$ – средние классов, а $\boldsymbol{\Sigma}^{\oplus}$ и $\boldsymbol{\Sigma}^{\ominus}$ – ковариационные матрицы для каждого класса. Чтобы лучше разобраться в этом, предположим, что $\boldsymbol{\Sigma}^{\oplus} = \boldsymbol{\Sigma}^{\ominus} = \mathbf{I}$ (то есть в каждом классе признаки некоррелированы и имеют единичную дисперсию), тогда имеем:

$$\mathrm{LR}(\mathbf{x}) = \exp\left(-\frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu}^{\oplus})^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}^{\oplus}) - (\mathbf{x}-\boldsymbol{\mu}^{\ominus})^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}^{\ominus})] \right) =$$

$$= \exp\left(-\frac{1}{2}[\|\mathbf{x}-\boldsymbol{\mu}^{\oplus}\|^2 - \|\mathbf{x}-\boldsymbol{\mu}^{\ominus}\|^2] \right).$$

Отсюда следует, что $\mathrm{LR}(\mathbf{x}) = 1$ для любого $\mathbf{x}$, равноудаленного от $\boldsymbol{\mu}^{\oplus}$ и $\boldsymbol{\mu}^{\ominus}$. Но это означает, что решающая граница с максимальным правдоподобием – прямая линия, проходящая на равном расстоянии от средних точек классов, – а это не

что иное, как наш старый приятель, базовый линейный классификатор! Другими словами, *для некоррелированных гауссовых признаков с единичной дисперсией базовый линейный классификатор является оптимальным по Байесу*. Это хороший пример того, как с помощью вероятностного подхода можно обосновать конкретные модели.

В общем случае при условии, что ковариационные матрицы для каждого класса одинаковы, решающая граница с максимальным правдоподобием будет гиперплоскостью, пересекающей отрезок $\mu^{\oplus} - \mu^{\ominus}$ в середине, но не под прямым углом, если признаки коррелированы. Это означает, что в этом случае базовый линейный классификатор будет оптимален по Байесу, только если предварительно устранить корреляцию и нормировать признаки. При неравных ковариационных матрицах классов решающая граница будет гиперболической. Таким образом, все три случая, показанные на рис. 9.3, обобщаются на многомерную ситуацию.

Мы видели несколько примеров того, как нормальное распределение связывает геометрический и вероятностный взгляды на вещи. Многомерное нормальное распределение по существу преобразует расстояния в вероятности. Это становится очевидным, если подставить определение ☞ *расстояния Махаланобиса* (формула (8.1) на стр. 248) в уравнение (9.2):

$$P(\mathbf{x}\,|\,\mu,\boldsymbol{\Sigma}) = \frac{1}{E_d}\exp\left(-\frac{1}{2}(\mathrm{Dis}_M(\mathbf{x},\mu\,|\,\boldsymbol{\Sigma}))^2\right). \tag{9.4}$$

Аналогично стандартное нормальное распределение преобразует евклидовы расстояния в вероятности:

$$P(\mathbf{x}\,|\,\mathbf{0},\mathbf{I}) = \frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{1}{2}(\mathrm{Dis}_2(\mathbf{x},\mathbf{0}))^2\right).$$

Обратно, мы видим, что *взятый со знаком минус логарифм гауссова правдоподобия можно интерпретировать как квадрат расстояния*:

$$-\ln P(\mathbf{x}\,|\,\mu,\boldsymbol{\Sigma}) = \ln E_d + \frac{1}{2}(\mathrm{Dis}_M(\mathbf{x},\mu\,|\,\boldsymbol{\Sigma}))^2.$$

Интуиция подсказывает, что логарифм преобразует мультипликативную шкалу вероятностей в аддитивную шкалу (которая в случае гауссова распределения соответствует квадрату расстояния). Поскольку с аддитивной шкалой работать проще, логарифмическое правдоподобие применяется в статистике очень часто.

Еще один пример связи между геометрическим и вероятностным взглядами возникает, когда мы рассматриваем вопрос об оценивании параметров нормального распределения. Например, предположим, что нужно оценить среднее μ многомерного гауссова распределения с заданной ковариационной матрицей $\boldsymbol{\Sigma}$, имея набор данных X. Принцип *оценки максимального правдоподобия* утверждает, что следует искать значение μ, которое обращает в максимум совместное правдоподобие X. В предположении, что элементы X выбирались независимо, совместное

правдоподобие разлагается в произведение по отдельным точкам X, и оценку максимального правдоподобия можно найти следующим образом:

$$\begin{aligned}
\hat{\mu} &= \arg\max_{\mu} \prod_{\mathbf{x}\in X} P(\mathbf{x}\,|\,\mu,\boldsymbol{\Sigma}) = \\
&= \arg\max_{\mu} \prod_{\mathbf{x}\in X} \frac{1}{E_d} \exp\left(-\frac{1}{2}(\mathrm{Dis}_M(\mathbf{x},\mu\,|\,\boldsymbol{\Sigma}))^2\right) = && \text{из уравнения (9.4)} \\
&= \arg\min_{\mu} \sum_{\mathbf{x}\in X} \left[\ln E_d + \frac{1}{2}(\mathrm{Dis}_M(\mathbf{x},\mu\,|\,\boldsymbol{\Sigma}))^2\right] = && \text{путем взятия логарифмов со знаком минус} \\
&= \arg\min_{\mu} \sum_{\mathbf{x}\in X} (\mathrm{Dis}_M(\mathbf{x},\mu\,|\,\boldsymbol{\Sigma}))^2. && \text{опускаем постоянный член и коэффициент}
\end{aligned}$$

Таким образом, находим, что оценка максимального правдоподобия среднего многомерного распределения – это точка, в которой обращается в минимум сумма квадратов расстояния Махаланобиса до всех точек X. В случае единичной ковариационной матрицы $\boldsymbol{\Sigma} = \mathbf{I}$ мы можем заменить расстояние Махаланобиса евклидовым, и тогда по теореме 8.1 точкой, минимизирующей сумму квадратов евклидовых расстояний до всех точек X, является среднее арифметическое $\frac{1}{|X|}\sum_{\mathbf{x}\in X}\mathbf{x}$.

И в качестве последнего примера того, как геометрический и вероятностный взгляды на одну и ту же проблему могут быть сильно связаны, я продемонстрирую вывод ☞ *решения задачи линейной регрессии по методу наименьших квадратов* (раздел 7.1) как оценки максимального правдоподобия. Для простоты обозначений будем рассматривать одномерный случай, обсуждавшийся в примере 7.1. Отправной точкой является гипотеза, что обучающие примеры (h_i, y_i) – это зашумленные измерения истинной функции $(x_i, f(x_i))$, то есть $y_i = f(x_i) + \epsilon_i$, где ϵ_i – независимые невязки с одинаковым распределением. (Обратите внимание на небольшое изменение обозначений – y_i больше не обозначает истинного значения функции.) Мы хотим вывести оценки максимального правдоподобия $\hat{y}_i$ значений $f(x_i)$. Их можно вывести, если предположить конкретное распределение шума, например гауссово с дисперсией σ^2. Тогда каждое y_i имеет нормальное распределение со средним $a + bx_i$ и дисперсией σ^2, а значит:

$$P(y_i\,|\,a,b,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-(a+bx_i))^2}{2\sigma^2}\right).$$

Поскольку описывающие шум члены ϵ_i независимы для различных i, то таковы же и y_i, и, следовательно, совместное распределение вероятностей по всем i – это просто произведение n гауссиан:

$$P(y_1,\ldots,y_n \mid a,b,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-(a+bx_i))^2}{2\sigma^2}\right) =$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\Sigma_{i=1}^{n}(y_i-(a+bx_i))^2}{2\sigma^2}\right).$$

Для простоты алгебраических манипуляций возьмем натуральный логарифм со знаком минус:

$$-\ln P(y_1,\ldots,y_n \mid a,b,\sigma^2) = \frac{n}{2}\ln 2\pi + \frac{n}{2}\ln\sigma^2 + \frac{\Sigma_{i=1}^{n}(y_i-(a+bx_i))^2}{2\sigma^2}.$$

Приравняв к нулю частные производные по a, b и σ^2 для нахождения максимума отрицательного логарифмического правдоподобия, мы получим три уравнения:

$$\sum_{i=1}^{n} y_i - (a+bx_i) = 0;$$
$$\sum_{i=1}^{n} (y_i - (a+bx_i))\,x_i = 0;$$
$$\frac{n}{2}\frac{1}{\sigma^2} - \frac{\Sigma_{i=1}^{n}(y_i-(a+bx_i))^2}{2(\sigma^2)^2}.$$

Первые два – по существу те же самые, что были выведены в примере 7.1, они дают $\hat{a} = \bar{y} - \hat{b}\bar{x}$ и $\hat{b} = \sigma_{xy}/\sigma_{xx}$ соответственно. Третье уравнение говорит, что сумма квадратов невязок равна $n\sigma^2$ и дает оценку максимального правдоподобия дисперсии шума в виде $(\Sigma_{i=1}^{n}(y_i - (a + bx_i))^2)/n$.

Обнадеживает, что вероятностный взгляд позволяет вывести решение задачи регрессии по (обычному) методу наименьших квадратов из чисто теоретических соображений. С другой стороны, полное рассмотрение потребовало бы учитывать также зашумленность значений x (обобщенный метод наименьших квадратов), но это усложняет математику, и решение может быть не единственным. Это показывает, что *хороший вероятностный подход к проблеме машинного обучения позволяет достичь баланса между солидным теоретическим основанием и прагматизмом, необходимым для получения рабочего решения*.

9.2 Вероятностные модели для категориальных данных

Чтобы чем-то занять себя во время долгих поездок на удаленное место отдыха, мы с сестрами часто развлекаемся играми, в которых участвуют проезжающие мимо машины. Например, просим друг друга отмечать машины определенного

цвета, из определенной страны или с определенной буквой в номере. Вопрос с двумя возможными ответами типа «является ли машина синей?» в статистике называется *испытанием Бернулли*. Такие испытания моделируются бинарной случайной величиной, для которой вероятность успеха фиксирована и одинакова в каждом независимом испытании. Мы использовали распределение Бернулли для моделирования события «почтовое сообщение хорошее» в примере 9.1. Над такой случайной величиной можно построить и другие распределения вероятности. Например, можно угадывать, сколько следующих машин будут синими: соответствующее распределение называется биномиальным. Или оценить, сколько проедет машин до появления первой машины из Голландии: это определение геометрического распределения. В замечании 9.2 приведены соответствующие определения.

Категориальные величины или признаки (их еще называют дискретными, или номинальными) в машинном обучении встречаются повсеместно. Пожалуй, самая распространенная форма распределения Бернулли моделирует вхождение слова в документ. То есть для i-го слова в словаре мы имеем случайную величину X_i с распределением Бернулли. Совместное распределение *битового вектора* $X = (X_1, ..., X_k)$ называется *многомерным распределением Бернулли*. Часто встречаются также величины с более чем двумя исходами: например, позиции слова в почтовом сообщении соответствует категориальная величина с k исходами, где k – размер словаря. Мультиномиальное распределение выступает в виде *вектора счетчиков*: гистограммы количества вхождений всех словарных слов в документ. Это дает альтернативный способ моделирования текстовых документов, позволяющий учитывать количество вхождений слова при классификации документа.

Распределение Бернулли, названное в честь швейцарского математика XVII века Якоба Бернулли, относится к булевым, или бинарным, событиям с двумя возможными исходами: успех (1) и неудача (0). У распределения Бернулли имеется единственный параметр θ, определяющий вероятность успеха: $P(X = 1) = \theta$ и $P(X = 0) = 1 - \theta$. Математическое ожидание распределения Бернулли $\mathbb{E}[X] = \theta$, а дисперсия $\mathbb{E}[(X - \mathbb{E}[X])^2] = \theta(1 - \theta)$.

Биномиальное распределение возникает при подсчете числа успехов S в n независимых испытаниях Бернулли с одним и тем же параметром θ. Оно описывается формулой

$$P(S = s) = \binom{n}{s}\theta^s(1-\theta)^{n-s} \text{ для } s \in \{0,\ldots,n\}.$$

Его математическое ожидание равно $\mathbb{E}[S] = n\theta$, а дисперсия $\mathbb{E}[(S - \mathbb{E}[S])^2] = n\theta(1 - \theta)$.

Категориальное распределение обобщает распределение Бернулли на $k \geq 2$ исходов. Его параметром является k-мерный вектор $\theta = (\theta_1, ..., \theta_k)$ – такой, что $\Sigma_{i=1}^k \theta_i = 1$.

Наконец, *мультиномиальное распределение* имеет исходы n независимых категориальных испытаний с одинаковым распределением. То есть $\mathbf{X} = (X_1, ..., X_k)$ – k-мерный вектор целочисленных счетчиков и

$$P(\mathbf{X} = (x_1,\ldots,x_k)) = n!\frac{\theta_1^{x_1}}{x_1!}\ldots\frac{\theta_k^{x_k}}{x_k!}$$

при условии $\Sigma_{i=1}^k x_i = n$. Отметим, что задание $n = 1$ дает альтернативный способ определения категориального распределения в виде $P(\mathbf{X} = (x_1, ..., x_k)) = \theta_1^{x_1} \cdots \theta_k^{x_k}$ при условии, что ровно один x_i

равен 1, а все остальные – 0. А задание $k = 2$ дает альтернативное выражение для распределения Бернулли в виде $P(X = x) = \theta^x(1 - \theta)^{1-x}$ при $x \in \{0,1\}$. Полезно также заметить, что если $\mathbf{X}$ имеет мультиномиальное распределение, то каждая компонента X_i имеет биномиальное распределение с параметром θ_i.

Мы можем оценить параметры этих распределений прямым подсчетом. Предположим, что *a b a c c b a a b c* – последовательность слов. Нас может интересовать, совпадает слово с *a* или нет, при том что данные интерпретируются как результаты независимых одинаково распределенных испытаний Бернулли; это даст оценку $\hat{\theta}_a = 4/10 = 0.4$. Тот же самый параметр порождает биномиальное распределение количества вхождений слова *a* в аналогичные последовательности. Точно так же можно оценить параметры категориального (вхождение слов) и мультиномиального (счетчики слов) в виде $\hat{\theta} = (0.4, 0.3, 0.3)$.

Почти всегда полезно сглаживать эти распределения, включая *псевдосчетчики*. Представьте, что словарь содержит слово *d*, но мы его еще не наблюдали, тогда оценка максимального правдоподобия имела бы вид $\hat{\theta}_d = 0$. Эту оценку можно сгладить, добавив виртуальное вхождение каждого слова в наблюдения, что дает $\hat{\theta}' = (5/14, 4/14, 4/14, 1/14)$. В случае биномиального распределения это не что иное, как поправка Лапласа.

Замечание 9.2. Распределения вероятности для категориальных данных

Обе эти модели документов широко употребляются. Несмотря на различия, в обеих предполагается независимость вхождений слов, что обычно называют *наивным байесовским предположением*. В мультиномиальной модели документа это следует из самого определения мультиномиального распределения, в котором предполагается, что слова в разных позициях независимо выбираются из одного и того же категориального распределения. В многомерной модели Бернулли мы предполагаем, что отдельные биты битового вектора статистически независимы, что позволяет вычислить совместную вероятность конкретного битового вектора $(x_1, ..., x_k)$ как произведение вероятностей его компонент $P(X_i = x_i)$. На практике предположение о независимости часто не выполняется: если мы знаем, что почтовое сообщение содержит слово «виагра», то можно с высокой степенью уверенности утверждать, что в нем встречается и слово «таблетка». Как бы то ни было, опыт показывает, что хотя наивное байесовское предположение почти всегда дает плохие оценки вероятностей, это зачастую не вредит качеству ранжирования. Это означает, что при условии разумного выбора порога классификации мы обычно получаем также и хорошую классификацию.

Использование наивной байесовской модели для классификации

Предположим, что мы выбрали одно из возможных распределений для моделирования данных X. В контексте классификации мы можем далее предположить, что это распределение зависит от класса, так что распределения $P(X|Y = \text{спам})$ и $P(X|Y = \text{неспам})$ различны. Чем сильнее они различаются, тем полезнее признаки X для классификации. Таким образом, для конкретного почтового сообщения мы вычисляем обе величины $P(X = x|Y = \text{спам})$ и $P(X = x|Y = \text{неспам})$ и применяем одно из нескольких возможных решающих правил:

- максимального правдоподобия (ML) – предсказать $\operatorname{argmax}_y P(X = x|Y = y)$;
- апостериорного максимума (MAP) – предсказать $\operatorname{argmax}_y P(X = x|Y = y) P(Y = y)$;
- откалиброванного правдоподобия – предсказать $\operatorname{argmax}_y w_y P(X = x|Y = y)$.

Первые два решающих правила соотносятся следующим образом: классификация по методу ML эквивалентна классификации по методу MAP, если распределение по классам равномерно. Третье правило обобщает первые два в том смысле, что заменяет распределение по классам набором весов, обученным на данных: как мы увидим ниже, это позволяет исправлять ошибки оценивания в правдоподобиях.

Пример 9.4 (предсказание с помощью наивной байесовской модели). Предположим, что словарь содержит три слова *a*, *b* и *c* и что для почтовых сообщений используется многомерная модель Бернулли с параметрами

$$\theta^{\oplus} = (0.5, 0.67, 0.33) \qquad \theta^{\ominus} = (0.67, 0.33, 0.33)$$

Это, в частности, означает, что присутствие *b* в два раза более вероятно в спаме (+), чем в хороших сообщениях.

Подлежащее классификации сообщение содержит слова *a* и *b*, но не *c*, и, следовательно, описывается битовым вектором $\mathbf{x} = (1,1,0)$. Получаются такие правдоподобия:

$$P(\mathbf{x}|\oplus) = 0.5 \cdot 0.67 \cdot (1–0.33) = 0.222 \qquad P(\mathbf{x}|\ominus) = 0.67 \cdot 0.33 \cdot (1–0.33) = 0.148$$

Следовательно, по правилу ML **x** классифицируется как спам. В случае двух классов часто удобнее работать с отношениями правдоподобия и шансами. Отношение правдоподобия можно вычислить как $\frac{P(\mathbf{x}|\oplus)}{P(\mathbf{x}|\ominus)} = \frac{0.5}{0.67}\frac{0.67}{0.33}\frac{1-0.33}{1-0.33} = 3/2 > 1$. Это означает, что по правилу MAP **x** также классифицируется как спам, если априорный шанс больше 2/3, и как неспам, если меньше. Например, при 33% спама и 67% неспама априорный шанс равен $\frac{P(\oplus)}{P(\ominus)} = \frac{0.33}{0.67} = 1/2$, а апостериорный, следовательно, $\frac{P(\oplus|\mathbf{x})}{P(\ominus|\mathbf{x})} = \frac{P(\mathbf{x}|\oplus)}{P(\mathbf{x}|\ominus)}\frac{P(\oplus)}{P(\ominus)} = 3/2 \cdot 1/2 = 3/4 < 1$. В этом случае отношение правдоподобия для **x** недостаточно велико, чтобы изменить решение, по сравнению с априорным.

Альтернативно можно использовать мультиномиальную модель. Ее параметры задают распределение слов из словаря, например:

$$\theta^{\oplus} = (0.3, 0.5, 0.2) \qquad \theta^{\ominus} = (0.6, 0.2, 0.2)$$

Подлежащее классификации сообщение содержит три вхождения слова *a*, одно вхождение слова *b* и не содержит слова *c*, то есть описывается вектором счетчиков $\mathbf{x} = (3,1,0)$. Общее число вхождений словарных слов равно $n = 4$. Получаются такие правдоподобия:

$$P(\mathbf{x}|\oplus) = 4!\frac{0.3^3}{3!}\frac{0.5^1}{1!}\frac{0.2^0}{0!} = 0.054; \qquad P(\mathbf{x}|\ominus) = 4!\frac{0.6^3}{3!}\frac{0.2^1}{1!}\frac{0.2^0}{0!} = 0.1728.$$

Отношение правдоподобия равно $\left(\frac{0.3}{0.6}\right)^3\left(\frac{0.5}{0.2}\right)^1\left(\frac{0.2}{0.2}\right)^0 = 5/16$. По правилу ML **x** классифицируется как неспам, то есть прямо противоположно многомерной модели Бернулли. Это случилось главным образом из-за трех вхождений слова *a*, что является сильным свидетельством в пользу неспама.

Отметим, что отношение правдоподобия для многомерной модели Бернулли является произведением сомножителей $\theta_i^{\oplus}/\theta_i^{\ominus}$, если $x_i = 1$ в классифицируемом битовом векторе и $(1 - \theta_i^{\oplus})/(1 - \theta_i^{\ominus})$, если $x_i = 0$. Для мультиномиальной модели сомножители равны $(\theta_i^{\oplus}/\theta_i^{\ominus})^{x_i}$. Из этого следует, в частности, что мультиномиальная модель принимает во внимание только присутствие слов, тогда как многомерная модель Бернулли – еще и их отсутствие. В примере выше отсутствию слова *b* соответствует сомножитель $(1 - 0.67)/(1 - 0.33) = 1/2$ в отношении правдоподобия. Еще одно существенное различие между этими двумя моделями заключается в том, что в мультиномиальной многократные вхождения слов трактуются как повторяющиеся признаки – благодаря «весу» x_i в показателе степени. Это становится понятнее, если взять логарифм отношения правдоподобия, равный $\sum_i x_i (\ln\theta_i^{\oplus} - \ln\theta_i^{\ominus})$: это выражение линейно зависит от $\ln\theta_i^{\oplus}$ и $\ln\theta_i^{\ominus}$ с весовыми коэффициентами x_i. Отметим, что это еще не означает линейность байесовских классификаторов в смысле, обсуждавшемся в главе 7, если только мы не сможем продемонстрировать линейное соотношение между $\ln\theta$ и значением соответствующего признака. Но мы можем сказать, что наивные байесовские модели линейны в специальном пространстве (пространстве «логарифмических шансов»), получающемся путем применения корректно определенного преобразования к признакам. Мы вернемся к этому моменту, когда будем обсуждать ☞ *калибровку признаков* в разделе 10.2.

Тот факт, что отношение совместного правдоподобия наивной байесовской модели разлагается в произведение отношений правдоподобия отдельных слов, – прямое следствие наивного байесовского предположения. Иными словами, задача обучения разлагается на одномерные задачи, по одной для каждого слова в словаре. Мы уже встречались с подобным разложением при обсуждении ☞ *многомерной линейной регрессии* в разделе 7.1. Там мы видели пример вреда, который может нанести игнорирование корреляции признаков. Можно ли привести аналогичные примеры для наивных байесовских классификаторов? Рассмотрим ситуацию, когда некоторое слово встречается в словаре дважды. В таком случае один и тот же сомножитель будет дважды входить в произведение, представляющее отношение правдоподобия, и по существу удваивает вес данного слова, по сравнению со всеми остальными. Хотя это крайний пример, такой двойной подсчет действительно дает заметный эффект на практике. Я уже выше указывал, что сообщение, содержащее слово «виагра», скорее всего, содержит и слово «таблетка», поэтому наблюдение обоих слов вместе не является существенно более сильным свидетельством в пользу спама, чем наблюдение одного лишь первого слова, а отношение правдоподобия для обоих слов не должно быть намного больше, чем для первого слова. Однако перемножение двух отношений правдоподобия, больших 1, дает величину, большую каждого из сомножителей. В результате оценки вероятностей, даваемые наивным байесовским классификатором, зачастую смещаются слишком далеко в сторону 0 или 1.

Если нас интересует только классификация, а не оценки вероятностей как таковые, то может показаться, что это не страшно. Однако *часто забывают о*

том, что для таких неоткалиброванных оценок вероятностей, какие порождаются наивным байесовским классификатором, решающие правила ML и MAP становятся неадекватными. Если у нас нет свидетельств, доказывающих, что предположения модели верны, то в таком случае единственная разумная вещь – воспользоваться *решающим правилом откалиброванного правдоподобия*, которое требует обучить вектор весов классов, чтобы исправить ошибки оценивания в правдоподобиях. Точнее, мы хотим найти веса w_i – такие, что предсказанная величина $\text{argmax}_y w_y P(X = x|Y = y)$ приводит к наименьшей возможной потере – числу неправильно классифицированных примеров – на тестовом наборе. Для двух классов эту задачу можно решить с помощью той же процедуры ☞ *преобразования ранжировщиков в классификаторы*, которую мы рассматривали в разделе 2.2. Чтобы убедиться в этом, заметим, что для двух классов решающее правило откалиброванного правдоподобия можно переписать в виде:

- ☞ предсказывать положительный класс, если $w^{\oplus}P(X = x|Y = \oplus) > w^{\ominus}P(X = x|Y = \ominus)$, и отрицательный в противном случае, что эквивалентно;
- ☞ предсказывать положительный класс, если $P(X = x|Y = \oplus)/P(X = x|Y = \ominus) > w^{\ominus}/w^{\oplus}$, и отрицательный в противном случае.

Это показывает, что в случае двух классов у нас на самом деле есть всего одна степень свободы, поскольку умножение весов на константу не влияет на принятие решений. Иными словами, нам интересно найти наилучший порог $t = w^{\ominus}/w^{\oplus}$ отношения правдоподобия, а это фактически та же задача, что и нахождение наилучшей рабочей точки на кривой РХП. Решением является точка на изолинии самой высокой верности. На рис. 9.4 изображены два реальных набора данных; слева мы видим, что порог принятия решения MAP более-менее оптимален, а справа – что оптимальная точка находится в правом верхнем углу.

Если классов больше двух, то объем вычислений, необходимых для нахождения глобально оптимального вектора весов, превышает наши возможности, а значит, необходимо прибегнуть к эвристическому методу. В разделе 3.1 такой метод был продемонстрирован для трех классов. Идея заключается в том, чтобы поочередно фиксировать веса, используя некоторое упорядочение классов. Иначе говоря, мы применяем двухклассовую процедуру, чтобы оптимально отделить i-й класс от предыдущих $i - 1$ классов.

Обучение наивной байесовской модели

Обучение вероятностной модели обычно сводится к оцениванию параметров, используемых в этой модели распределений. Параметр распределения Бернулли можно оценить, подсчитав число успехов d в n испытаниях и положив $\hat{\theta} = d/n$. Иными словами, мы для каждого класса подсчитываем, сколько сообщений содержит рассматриваемое слово. Такие оценки на основе относительной частоты обычно сглаживаются путем введения *псевдосчетчиков*, представляющих исходы виртуальных испытаний с некоторыми фиксированными распределениями. В случае распределения Бернулли в качестве операции сглаживания чаще все-

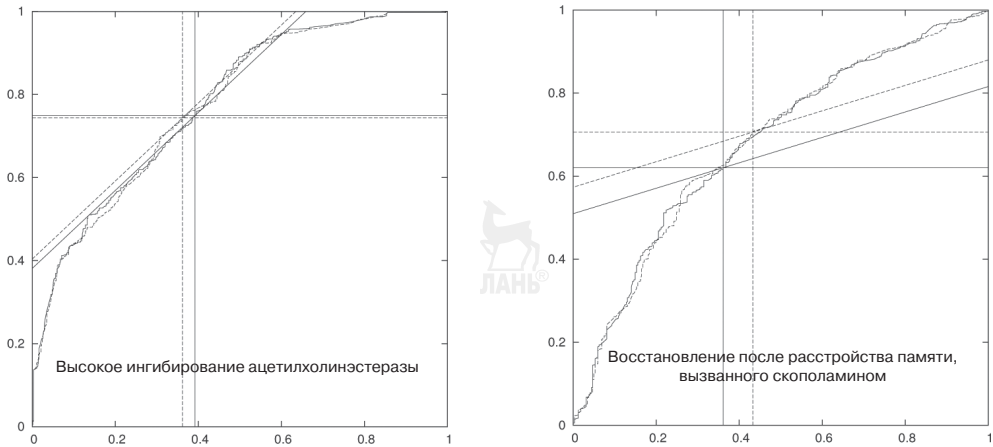

Рис. 9.4. (Слева) Кривые РХП, порожденные двумя наивными байесовскими классификаторами (сплошная линия: вариант многомерной модели Бернулли; штриховая линия: вариант мультиномиальной модели). У обеих моделей схожее качество ранжирования и почти одинаковый – более-менее оптимальный – порог принятия решения MAP. **(Справа)** На другом наборе данных из той же предметной области порог MAP для мультиномиальной модели несколько лучше, что наводит на мысль о лучших калиброванных оценках вероятностей. Но поскольку угловой коэффициент изолиний верность показывает, что на каждый отрицательный пример приходится примерно четыре положительных, то оптимальное решающее правило фактически всегда будет предсказывать положительный класс

го берут поправку Лапласа, которая подразумевает два виртуальных испытания: одно успешное, другое неудачное. Следовательно, оценка на основе относительной частоты заменяется на $(d + 1)/(n + 2)$. С точки зрения байесовской модели, это сводится к принятию равномерного априорного распределения, выражающего нашу веру в том, что успех и неудача равновероятны. Если того требует ситуация, мы можем усилить влияние априорного знания, включив большее количество виртуальных испытаний, что означает, что для отодвигания оценки от априорной потребуется больше данных. В случае категориального распределения сглаживание добавляет один псевдосчетчик для каждой из k категорий, что приводит к сглаженной оценке $(d + 1)/(n + k)$. *m-оценка* представляет собой дальнейшее обобщение, считая параметрами как общее число псевдосчетчиков m, так и их распределение по категориям. Оценка для i-й категории определяется как $(d + p_i m)/(n + m)$, где p_i – распределение по категориям (то есть $\sum_{i=1}^{k} p_i = 1$). Отметим, что оценки на основе сглаженной относительной частоты – а значит, и произведение таких оценок – никогда не могут достигать крайних значений $\hat{\theta} = 0$ и $\hat{\theta} = 1$.

Пример 9.5 (обучение наивной байесовской модели). Сейчас мы покажем, как можно было бы получить векторы параметров в предыдущем примере. Рассмотрим следующие сообщения, состоящие из пяти слов a, b, c, d, e:

e_1: *b d e b b d e*
e_2: *b c e b b d d e c c*
e_3: *a d a d e a e e*
e_4: *b a d b e d a b*
e_5: *a b a b a b a e d*
e_6: *a c a c a c a e d*
e_7: *e a e d a e a*
e_8: *d e d e d*

Нам известно, что сообщения слева – спам, а справа – неспам, и мы хотим использовать их в качестве небольшого обучающего набора для обучения байесовского классификатора. Сначала мы принимаем решение, что *d* и *e* – так называемые *стоп-слова*, которые встречаются настолько часто, что не несут никакой информации о классе. Остальные слова, *a*, *b* и *c*, составляют наш словарь.

В случае мультиномиальной модели мы представляем каждое сообщение вектором счетчиков, как в табл. 9.1 слева. Чтобы оценить параметры распределения, мы вычисляем сумму векторов счетчиков для каждого класса и получаем (5, 9, 3) для спама и (11, 3, 3) для неспама. Чтобы сгладить эти оценки вероятностей, мы добавляем по одному псевдосчетчику для каждого словарного слова, что доводит общее число вхождений словарных слов до 20 для каждого класса. Таким образом, оценочные векторы параметров равны $\hat{\theta}^{\oplus} = (6/20, 10/20, 4/20) = (0.3, 0.5, 0.2)$ для спама и $\hat{\theta}^{\ominus} = (12/20, 4/20, 4/20) = (0.6, 0.2, 0.2)$ для неспама.

В случае многомерной модели Бернулли сообщения представлены битовыми векторами, как в табл. 9.1 справа. Сложение битовых векторов для каждого класса дает (2, 3, 1) спама и (3, 1, 1) для неспама. Каждый счетчик следует разделить на количество документов в классе, чтобы получить оценку вероятности того, что документ содержит конкретное словарное слово. Сглаживание вероятности теперь означает добавление двух псевдодокументов, один из которых содержит все слова, а другой – ни одного. Это дает такие оценочные векторы параметров: $\hat{\theta}^{\oplus} = (3/6, 4/6, 2/6) = (0.5, 0.67, 0.33)$ для спама и $\hat{\theta}^{\ominus} = (4/6, 2/6, 2/6) = (0.67, 0.33, 0.33)$ для неспама.

Сообщение	*#a*	*#b*	*#c*	*Класс*
e_1	0	3	0	+
e_2	0	3	3	+
e_3	3	0	0	+
e_4	2	3	0	+
e_5	4	3	0	–
e_6	4	0	3	–
e_7	3	0	0	–
e_8	0	0	0	–

Сообщение	*a?*	*b?*	*c?*	*Класс*
e_1	0	1	0	+
e_2	0	1	1	+
e_3	1	0	0	+
e_4	1	1	0	+
e_5	1	1	0	–
e_6	1	0	1	–
e_7	1	0	0	–
e_8	0	0	0	–

Таблица 9.1. (Слева) Небольшой набор данных о почтовых сообщениях, описываемый векторами счетчиков. **(Справа)** Тот же самый набор данных, описываемый битовыми векторами

Существует много других вариантов наивного байесовского классификатора. На самом деле в том, что обычно понимается под «настоящим» байесовским классификатором, используется не мультиномиальная модель и не многомерная модель Бернулли, а многомерная категориальная модель. Это означает, что все признаки категориальные и что вероятность того, что *i*-й признак принимает свое *l*-ое значение для примеров из класса *c*, равна $\theta_{il}^{(c)}$ при условии, что $\sum_{l=1}^{k_i} \theta_{il}^{(c)} = 1$, где k_i – количество различных значений *i*-го признака. Эти параметры можно оценить с помощью сглаженных относительных частот в обучающем наборе, как

и в случае многомерной модели Бернулли. Совместная вероятность вектора признаков снова равна произведению вероятностей отдельных признаков, а значит, $P(F_i, F_j|C) = P(F_i|C)P(F_j|C)$ для всех пар признаков и для всех классов.

Отметим попутно, что условная независимость не имеет ничего общего с безусловной: ни одна не следует из другой. Чтобы убедиться в том, что из условной независимости не вытекает безусловная, представим два слова, появление которых в спаме очень вероятно, но при этом они независимы (то есть вероятность их одновременного появления в спамном сообщении равна произведению маргинальных вероятностей). Предположим также, что их появление в хорошем сообщении крайне маловероятно, хотя эти события также независимы. Допустим, я говорю вам, что неклассифицированное сообщение содержит одно из этих слов; надо думать, вы решите, что сообщение спамное, а отсюда сделаете вывод, что оно содержит и другое слово, – демонстрация того, что слова не являются безусловно независимыми. Чтобы понять, почему безусловная независимость не влечет за собой условной, рассмотрим два разных независимых слова, и пусть сообщение является спамом, если оно содержит хотя бы одно из этих слов, и неспамом в противном случае. Тогда среди спамных сообщений оба слова зависимы (поскольку если я знаю, что спамное сообщение не содержит одного из них, то должно содержать другое).

Другое обобщение наивной байесовской модели необходимо, когда некоторые признаки принимают вещественные значения. Одна из возможностей – дискретизировать их на стадии предобработки – обсуждается в главе 10. Другая – предположить, что значения признаков имеют нормальное распределение в каждом классе, как обсуждалось в предыдущем разделе. В этом контексте стоит отметить, что наивное байесовское предположение сводится к предположению о диагональности ковариационной матрицы в каждом классе, так что все признаки можно рассматривать независимо. Третья возможность, которая также применяется на практике, – смоделировать условное относительно класса правдоподобие каждого признака с помощью непараметрической оценки плотности. Все три варианта показаны на рис. 9.5.

Короче говоря, наивная байесовская модель популярна при работе с текстовыми, категориальными и смешанными – категориальными и вещественными – данными. Ее основной недостаток в качестве вероятностной модели – плохо откалиброванные оценки вероятностей – перевешивается в общем неплохим качеством ранжирования. Еще один очевидный парадокс наивной байесовской модели – то, что она вовсе и не байесовская! Во-первых, мы видели, что плохое качество оценок вероятности вынуждает использовать взвешенные по-другому правдоподобия, то есть вообще отказываться от применения правила Байеса. Во-вторых, при обучении наивной байесовской модели мы оцениваем параметры методом максимального правдоподобия, тогда как в настоящем байесовском подходе мы не интересуемся значениями отдельных параметров, а используем полное апостериорное распределение. Лично мне кажется, что смысл наивной байесовской модели заключается в разложении совместных правдоподобий

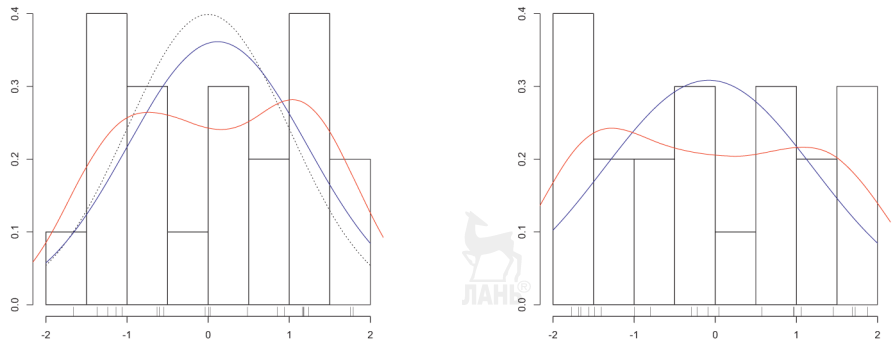

Рис. 9.5. (Слева) Примеры трех оценок плотности на выборке из 20 точек, имеющей нормальное распределение с нулевым средним и единичной дисперсией (штриховая линия). Гистограмма представляет собой простой непараметрический метод, в котором используется фиксированное число интервалов равной длины. Ядерная оценка плотности (**красная** линия) получается применением интерполяции для сглаживания функции плотности. Сплошная колоколообразная кривая (**синяя**) получена путем оценивания выборочного среднего и дисперсии в предположении, что истинное распределение нормально. **(Справа)** Здесь 20 точек распределены равномерно на отрезке [–2, 2], и непараметрические методы в общем случае работают лучше

в произведение маргинальных. Это разложение очень выразительно представлено в виде шотландского пледа в клетку на рис. 1.3, поэтому я называю наивный байесовский классификатор «шотландским».

9.3 Дискриминантное обучение путем оптимизации условного правдоподобия

Во введении к этой главе мы провели различие между порождающими и дискриминантными вероятностными моделями. Наивные байесовские модели порождающие: после обучения их можно использовать для порождения новых данных. В этом разделе мы рассмотрим наиболее распространенную дискриминантную модель: *логистическую регрессию*[1]. Проще всего понять логистическую регрессию, рассматривая ее как линейный классификатор, оценки вероятностей которого логистически калиброваны методом, описанным в разделе 7.4, но с одним существенным отличием: калибровка является неотъемлемой частью алгоритма обучения, а не шагом постобработки. Если в порождающих моделях решающая граница – побочный продукт моделирования распределений каждого класса, то логистическая регрессия моделирует решающую границу непосредственно. На-

[1] Отметим, что слово «регрессия» здесь не вполне уместно, поскольку, несмотря на то что оценка вероятности аппроксимирует неизвестную функцию, обучающие метки являются классами, а не значениями функции.

пример, если классы перекрываются, это значит, что логистическая регрессия демонстрирует тенденцию располагать решающую границу в области, где перекрытие классов максимально, вне зависимости от «форм» выборок из каждого класса. Получающие в результате решающие границы заметно отличаются от тех, что строят обученные порождающие классификаторы (рис. 9.6).

Уравнение (7.13) на стр. 233 выражает отношение правдоподобия в виде $\exp(\gamma(d(\mathbf{x}) - d_0))$, где $d(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - t$. Так как в случае дискриминантного обучения мы обучаем все параметры сразу, то γ и d_0 можно включить в $\mathbf{w}$ и t. Поэтому модель логистической регрессии описывается просто уравнением:

$$\hat{p}(\mathbf{x}) = \frac{\exp(\mathbf{w} \cdot \mathbf{x} - t)}{\exp(\mathbf{w} \cdot \mathbf{x} - t) + 1} = \frac{1}{1 + \exp(-(\mathbf{w} \cdot \mathbf{x} - t))}.$$

В предположении, что метки классов $y = 1$ для положительных примеров и $y = 0$ для отрицательных, эта формула определяет распределение Бернулли для каждого обучающего примера:

$$P(y_i | \mathbf{x}_i) = \hat{p}(\mathbf{x}_i)^{y_i}(1 - \hat{p}(\mathbf{x}_i))^{(1-y_i)}.$$

Важно отметить, что параметры этих распределений Бернулли связаны между собой посредством $\mathbf{w}$ и t, и, следовательно, существует один параметр для каждого признака, а не по одному для каждого обучающего примера.

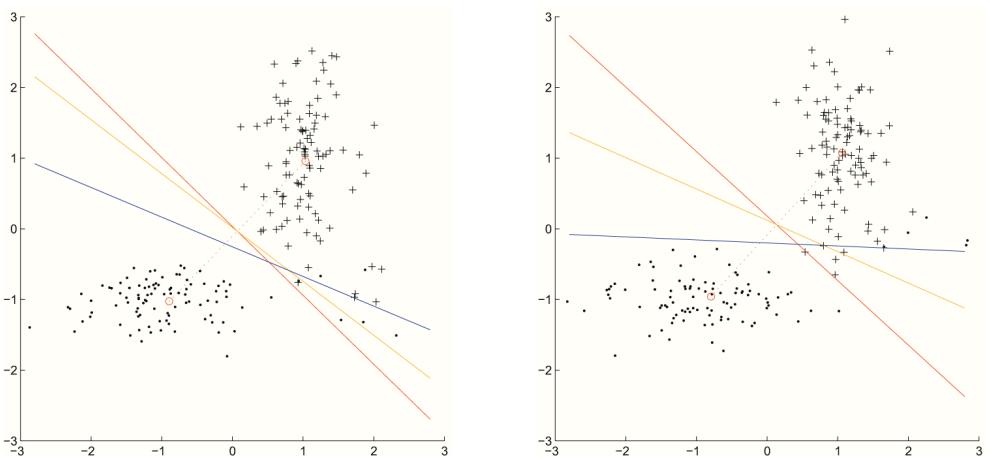

Рис. 9.6. (Слева) На этом наборе данных логистическая регрессия (**синяя** линия) лучше базового линейного классификатора (**красная** линия) и классификатора по методу наименьших квадратов (**оранжевая** линия), потому что последние два более чувствительны к форме классов, тогда как логистическая регрессия сосредоточена на области, где классы перекрываются. **(Справа)** На этом, лишь немного отличающемся наборе данных оба метода оказываются лучше логистической регрессии, потому что она чрезмерно концентрируется на отслеживании перехода от преимущественно положительной к преимущественно отрицательной области

Функция правдоподобия имеет вид:

$$\mathrm{CL}(\mathbf{w},t)=\prod_i P(y_i\,|\,\mathbf{x}_i)=\prod_i \hat{p}(\mathbf{x}_i)^{y_i}(1-\hat{p}(\mathbf{x}_i))^{(1-y_i)}.$$

Она называется *условным правдоподобием*, чтобы подчеркнуть, что дает *условную* вероятность $P(y_i|\,\mathbf{x}_i)$, а не $P(\mathbf{x}_i)$, как в порождающей модели. Отметим, что для использования произведения необходимо предположение о независимости значений y при заданном $\mathbf{x}$; но это абсолютно разумное предположение, далеко не такое сильное, как наивное байесовское предположение о независимости $\mathbf{x}$ в каждом классе. Как обычно, проще работать с логарифмом функции правдоподобия:

$$\mathrm{LCL}(\mathbf{w},t)=\sum_i y_i \ln\hat{p}(\mathbf{x}_i)+(1-y_i)\ln(1-\hat{p}(\mathbf{x}_i))=$$
$$=\sum_{\mathbf{x}^{\oplus}\in Tr^{\oplus}}\ln\hat{p}(\mathbf{x}^{\oplus})+\sum_{\mathbf{x}^{\ominus}\in Tr^{\ominus}}\ln(1-\hat{p}(\mathbf{x}^{\ominus})).$$

Мы хотим максимизировать логарифмическое условное правдоподобие относительно этих параметров, а это означает, что все частные производные должны быть равны нулю:

$$\nabla_{\mathbf{w}}\,\mathrm{LCL}(\mathbf{w},t)=\mathbf{0};$$
$$\frac{\partial}{\partial t}\mathrm{LCL}(\mathbf{w},t)=0.$$

Хотя эти уравнения не дают аналитического решения, их можно использовать для лучшего понимания природы логистической регрессии. Сосредоточимся на t и проделаем сначала подготовительные алгебраические преобразования:

$$\ln\hat{p}(\mathbf{x})=\ln\frac{\exp(\mathbf{w}\cdot\mathbf{x}-t)}{\exp(\mathbf{w}\cdot\mathbf{x}-t)+1}=\mathbf{w}\cdot\mathbf{x}-t-\ln(\exp(\mathbf{w}\cdot\mathbf{x}-t)+1);$$
$$\frac{\partial}{\partial t}\ln\hat{p}(\mathbf{x})=-1-\frac{\partial}{\partial t}\ln(\exp(\mathbf{w}\cdot\mathbf{x}-t)+1)=$$
$$=-1-\frac{1}{\exp(\mathbf{w}\cdot\mathbf{x}-t)+1}\exp(\mathbf{w}\cdot\mathbf{x}-t)\cdot(-1)=\hat{p}(\mathbf{x})-1.$$

Аналогично для отрицательных примеров:

$$\ln(1-\hat{p}(\mathbf{x}))=\ln\frac{1}{\exp(\mathbf{w}\cdot\mathbf{x}-t)+1}=-\ln(\exp(\mathbf{w}\cdot\mathbf{x}-t)+1);$$
$$\frac{\partial}{\partial t}\ln(1-\hat{p}(\mathbf{x}))=\frac{\partial}{\partial t}-\ln(\exp(\mathbf{w}\cdot\mathbf{x}-t)+1)=$$
$$=\frac{-1}{\exp(\mathbf{w}\cdot\mathbf{x}-t)+1}\exp(\mathbf{w}\cdot\mathbf{x}-t)\cdot(-1)=\hat{p}(\mathbf{x}).$$

Отсюда следует, что частная производная LCL по t записывается просто:

$$\frac{\partial}{\partial t}\mathrm{LCL}(\mathbf{w},t)=\sum_{\mathbf{x}^{\oplus}\in Tr^{\oplus}}(\hat{p}(\mathbf{x})-1)+\sum_{\mathbf{x}^{\ominus}\in Tr^{\ominus}}\hat{p}(\mathbf{x}^{\ominus})=\sum_{\mathbf{x}_i\in Tr}(\hat{p}(\mathbf{x}_i)-y_i).$$

Для оптимального решения эта частная производная равна нулю. Это означает, что в среднем предсказанная вероятность должна быть равна доле положительных примеров *pos*. Это удовлетворительный результат, поскольку это, очевидно, желательное глобальное свойство калиброванного классификатора.

Отметим, что группирующие модели, например деревья оценивания вероятностей, обладают этим свойством по построению, поскольку предсказанная вероятность в них считается равной эмпирической вероятности в сегменте.

Очень похожее рассуждение ведет к частной производной логарифмической условной вероятности по j-му весу w_j. Здесь следует отметить, что в то время как $\frac{\partial}{\partial t}(\mathbf{w}\cdot\mathbf{x}-t)=-1$, мы имеем $\frac{\partial}{\partial w_j}(\mathbf{w}\cdot\mathbf{x}-t)=\frac{\partial}{\partial w_j}(\Sigma_j w_j x_j-t)=x_j$ – j-му признаку объекта. Отсюда

$$\frac{\partial}{\partial t}\mathrm{LCL}(\mathbf{w},t)=\sum_{\mathbf{x}_i\in Tr}(\mathrm{y}_i-\hat{p}(\mathbf{x}_i))\mathrm{x}_{ij}. \qquad (9.5)$$

Приравнивание этой частной производной к нулю выражает еще одно калибровочное свойство на уровне отдельных признаков. Например, если j-ый признак является булевым и разреженным, преимущественно равным нулю, то это калибровочное свойство включает только объекты $\mathbf{x}i$, для которых $xij = 1$: в среднем предсказанная вероятность для этих объектов должна быть равна доле положительных среди них.

Пример 9.6 (одномерная логистическая регрессия). Рассмотрим данные на рис. 9.7, по 20 точек в каждом классе. Хотя оба класса были выбраны из нормальных распределений, перекрытие классов в данном примере меньше, чем можно было бы ожидать на основе их средних. Логистическая регрессия может воспользоваться этим фактом и дает гораздо более крутой сигмоид, чем базовый линейный классификатор с логистической калибровкой (см. объяснение в примере 7.7 на стр. 233), который формулируется всецело в терминах средних и дисперсий классов. Показаны также оценки вероятностей, полученные из выпуклой оболочки кривой РХП (см. рис. 7.13 на стр. 235); эта процедура калибровки непараметрическая и потому способна лучше обнаружить ограниченное перекрытие классов.

В терминах статистики логистическая регрессия имеет лучшую среднеквадратичную ошибку (0.040), чем логистически калиброванный классификатор (0.057). Изотонная калибровка дает наименьшую ошибку (0.021), но отметим, что не применялось сглаживание вероятностей для снижения риска переобучения. Сумма предсказанных вероятностей равна 18.7 для логистически калиброванного классификатора и 20 для двух других, то есть равна числу примеров, что является необходимым условием полной калибровки. Наконец, $\sum_{\mathbf{x}_i\in Tr}(y_i-\hat{p}(\mathbf{x}_i))x_i$ равна 2.6 для логистически калиброванного классификатора, 4.7 – для РХП-калиброванного классификатора и 0 – для логистической регрессии, чего и следовало ожидать из уравнения (9.5).

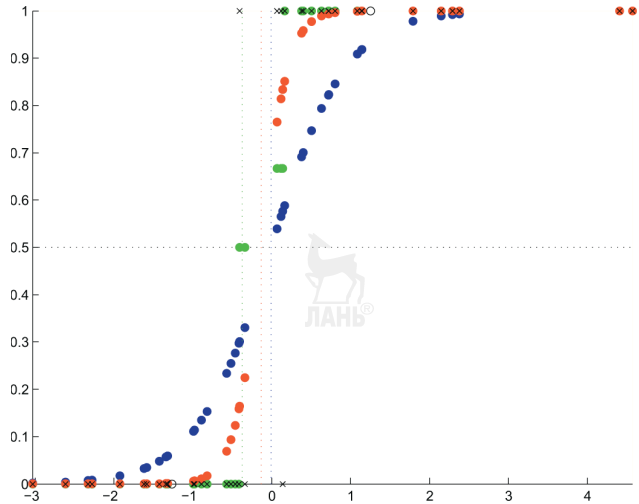

Рис. 9.7. Логистическая регрессия (**красная** линия) в сравнении с оценками вероятностей, полученными логистической калибровкой (**синяя** линия) и изотонной калибровкой (**зеленая** линия); последние две применены к базовому линейному классификатору (оценочные средние точки классов обозначены кружочками). Три соответствующие решающие границы показаны вертикальными пунктирными линиями

Для обучения логистической регрессионной модели нам нужно найти

$$\mathbf{w}^*, t^* = \underset{\mathbf{w},t}{\operatorname{argmax}} \mathrm{CL}(\mathbf{w},t) = \underset{\mathbf{w},t}{\operatorname{argmax}} \mathrm{LCL}(\mathbf{w},t).$$

Можно доказать, что это выпуклая задача оптимизации, а значит, у нее есть всего один максимум. Существует целый ряд методов ее решения. Простой подход, основанный на алгоритме обучения перцептрона, предполагает перебор примеров с использованием следующего правила обновления:

$$\mathbf{w} = \mathbf{w} + \eta(y_i - \hat{p}_i)\mathbf{x}_i,$$

где η – скорость обучения. Обратите внимание на связь с частной производной в уравнении (9.5). По существу, мы используем одиночные примеры для аппроксимации направления наискорейшего подъема.

9.4 Вероятностные модели со скрытыми переменными

Предположим, что перед нами стоит четырехклассовая задача классификации с классами *A*, *B*, *C* и *D*. При наличии достаточно объемной и репрезентативной обучающей выборки размера *n* мы можем использовать относительные частоты

в выборке n_A, ..., n_D для оценивания априорного распределения по классам $\hat{p}_A = n_A/n$, ..., $\hat{p}_D = n_D/n$, как уже много раз делали прежде[1]. Обратно, если известно априорное распределение и нужно узнать наиболее вероятное распределение по классам в случайной выборке из n объектов, то мы могли бы воспользоваться априорным распределением для вычисления математических ожиданий $\mathbb{E}[n_A] = p_A \cdot n$, ..., $\mathbb{E}[n_D] = p_D \cdot n$. Таким образом, полное знание одного позволяет оценить или вывести другое. Однако иногда мы обладаем лишь частичными знаниями о том и другом. Например, мы можем знать, что $p_A = 1/2$ и что C в два раза вероятнее B, не зная всего априорного распределения. И еще мы можем знать, что выборка, которую мы видели на прошлой неделе, была поровну разнесена между $A \cup B$ и $C \cup D$ и что C и D имели одинаковый размер, но не можем вспомнить размеров A и B порознь. И что теперь делать?

Формализуя все, что нам известно об априорном распределении, мы имеем: $p_A = 1/2$; $p_B = \beta$ (пока неизвестна); $p_C = 2\beta$, поскольку она в два раза больше p_B, и $p_D = 1/2 - 3\beta$, поскольку сумма всех четырех вероятностей должна быть равна 1. Далее: $n_A + n_B = a + b = s$, $n_C = c$ и $n_D = d$, где s, c и d известны. Требуется вычислить a, b и β, однако мы, похоже, столкнулись с проблемой курицы и яйца. Если бы мы знали β, то располагали бы полным знанием об априорном распределении и могли бы вывести математические ожидания a и b:

$$\frac{\mathbb{E}[a]}{\mathbb{E}[b]} = \frac{1/2}{\beta}; \quad \mathbb{E}[a] + \mathbb{E}[b] = s,$$

откуда

$$\mathbb{E}[a] = \frac{1}{1+2\beta} s; \quad \mathbb{E}[b] = \frac{2\beta}{1+2\beta} s. \tag{9.6}$$

Так, например, если $s = 20$ и $\beta = 1/10$, то $\mathbb{E}[a] = 16\frac{2}{3}$ и $\mathbb{E}[b] = 3\frac{1}{3}$.

Обратно, зная a и b, мы могли бы оценить β, применив оценку максимального правдоподобия и используя мультиномиальное распределение a, b, c и d:

$$P(a,b,c,d|\beta) = K(1/2)^a \beta^b (2\beta)^c (1/2 - 3\beta)^d,$$
$$\ln P(a,b,c,d|\beta) = \ln K + a \ln(1/2) + b \ln\beta + c \ln(2\beta) + d \ln(1/2 - 3\beta).$$

Здесь K – комбинаторная постоянная, которая не влияет на значение β, доставляющее максимум правдоподобию. Частная производная по β равна

$$\frac{\partial}{\partial\beta} \ln P(a,b,c,d\,|\,\beta) = \frac{b}{\beta} + \frac{2c}{2\beta} - \frac{3d}{1/2 - 3\beta}.$$

---

[1] Разумеется, если вы не уверены, достаточно ли велика выборка, то лучше сгладить эти оценки на основе относительных частот, например с помощью ☞ *поправки Лапласа* (раздел 2.3).

Приравнивая ее к нулю и решая уравнение относительно β, мы, наконец, получаем:

$$\hat{\beta} = \frac{b+c}{6(b+c+d)}. \tag{9.7}$$

Например, если $b = 5$ и $c = d = 10$, то $\hat{\beta} = 1/10$.

Чтобы разорвать этот замкнутый круг, мы можем повторять следующие два шага: (*i*) вычислить математическое ожидание отсутствующих частот a и b по предполагаемому или ранее оцененному значению параметра β; (*ii*) вычислить оценку максимального правдоподобия параметра β по предполагаемым или ранее оцененным математическим ожиданиям отсутствующих частот a и b. Эти два шага повторяются до достижения стационарного состояния. Так, если начать с $a = 15$, $b = 5$ и $c = d = 10$, то, как мы только что видели, $\hat{\beta} = 1/10$. Подставляя это значение β в уравнение (9.6), получаем $\mathbb{E}[a] = 16\frac{2}{3}$ и $\mathbb{E}[b] = 3\frac{1}{3}$. Подставляя эти значения обратно в уравнение (9.7), получаем $\hat{\beta} = 2/21$, что, в свою очередь, дает $\mathbb{E}[a] = 16.8$ и $\mathbb{E}[b] = 3.2$ и т. д. Для достижения стационарного состояния, в котором $\beta = 0.0948$, $a = 16.813$ и $b = 3.187$, требуется менее 10 итераций. В данном простом случае это действительно глобальный оптимум, который достигается независимо от начальной точки просто потому, что связь между b и β монотонна ($\mathbb{E}[b]$ возрастает вместе с β, согласно уравнению (9.6), а $\hat{\beta}$ возрастает вместе с b, согласно уравнению (9.7)). Однако в общем случае это не так, мы еще вернемся к этому моменту ниже.

EM-алгоритм

Проблема, которую мы только что обсуждали, – пример задач с отсутствующими данными, когда все множество данных Y разбивается на наблюдаемые переменные X и *скрытые переменные* Z (иногда их называют еще *латентными переменными*). В нашем примере наблюдаемыми были переменные c, d и s, а скрытыми – переменные a и b. Имеются также параметры модели θ – в нашем случае один параметр β[1]. Обозначим оценку θ на t-ой итерации θ^t. Существуют две относящиеся к делу величины:

- математическое ожидание $\mathbb{E}[Z|X,\theta^t]$ скрытой переменной при условии наблюдаемых переменных и текущей оценки параметров (так, в уравнении (9.6) математические ожидания a и b зависят от s и β);
- правдоподобие $P(Y|\theta)$, которое используется для нахождения максимизирующего значения θ.

[1] Параметры модели также в некотором роде «скрыты», но отличаются от скрытых переменных тем, что мы и не рассчитываем наблюдать их значения (например, средней точки класса), тогда как скрытую переменную, в принципе, можно было бы наблюдать, но вот в данном конкретном случае с этим не сложилось.

В функции правдоподобия нам нужны значения $Y = X \cup Z$. Очевидно, что мы используем наблюдаемые значения X, но для Z нам необходимы ранее вычисленные математические ожидания. Это означает, что на самом деле мы хотим максимизировать величину $P(X \cup \mathbb{E}[Z|X,\theta^t] \mid \theta)$ или, что эквивалентно, ее логарифм. Сделаем теперь предположение, что логарифм функции правдоподобия линейно зависит от Y: отметим, что в рассмотренном выше примере это предположение выполнялось. Для любой линейной функции f справедливо равенство $f(\mathbb{E}[Z]) = \mathbb{E}[f(Z)]$, и, следовательно, мы можем вынести математическое ожидание из целевой функции:

$$\ln P(X \cup \mathbb{E}[Z|X,\theta^t]|\theta) = \mathbb{E}[\ln P(X \cup Z|\theta)|X,\theta^t] = \mathbb{E}[\ln P(Y|\theta)|X,\theta^t]. \qquad (9.8)$$

Это последнее выражение обычно обозначают $Q(\theta|\theta^t)$, поскольку по существу оно говорит, как вычислить следующее значение θ, зная текущее:

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^t) = \arg\max_{\theta} \mathbb{E}[\ln P(Y|\theta)|X,\theta^t]. \qquad (9.9)$$

Это и есть общая форма знаменитого алгоритма *Expectation-Maximisation* (*EM-алгоритма*) – эффективного подхода к построению вероятностных моделей со скрытыми переменными или отсутствующими данными. Как в рассмотренном выше примере, мы производим итерации, каждая из которых состоит из двух шагов: присваивание ожидаемых значений скрытым переменным при условии текущих оценок параметров и переоценка параметров с учетом обновленных ожидаемых значений – и так до тех пор, пока не будет достигнуто стационарное состояние. Чтобы начать итерации, мы можем каким-то образом инициализировать либо параметры, либо скрытые переменные. Этот алгоритм очень напоминает алгоритм ☞ *K средних* (алгоритм 8.1 на стр. 260), в котором также производятся итерации из двух шагов: назначение точек в качестве текущих средних кластеров и переоценка средних кластеров с учетом новых назначений. Вскоре мы увидим, что это сходство не случайно. Как и в случае алгоритма K средних, можно доказать, что EM-алгоритм сходится к стационарному состоянию для широкого класса вероятностных моделей. Однако EM-алгоритм может остановиться в локальном минимуме при неудачном выборе начальной конфигурации.

Гауссовы смесовые модели

Хорошо известное применение EM-алгоритма – оценивание параметров *гауссовой смесовой модели* по данным. В такой модели данные берутся из K нормальных распределений, каждое со своим средним μ_j и ковариационной матрицей $\mathbf{\Sigma}_j$, причем доля точек, взятых из каждой гауссианы, определяется априорным распределением $\tau = (\tau_1, ..., \tau_K)$. Если бы каждая точка в выборке была помечена индексом гауссианы, из которой она взята, то мы имели бы простую задачу классификации, которую было бы легко решить, независимо оценив μ_j и $\mathbf{\Sigma}_j$ каждой гауссианы по данным, принадлежащим классу j. Однако мы сейчас рассматриваем гораздо

более трудную задачу прогностической кластеризации, в которой метки классов скрыты и должны быть восстановлены по наблюдаемым значениям признаков.

Для построения соответствующей модели удобно для каждой точки $\mathbf{x}_i$ иметь битовый вектор $\mathbf{z}_i = (z_{i1}, ..., z_{iK})$, в котором ровно один бит z_{ij} равен 1, а все остальные – 0. Такой вектор говорит о том, что i-я точка взята из j-ой гауссианы. В этих обозначениях мы можем модифицировать формулу ☞ *многомерного нормального распределения* (уравнение (9.2)) для получения общего выражения для гауссовой смесовой модели:

$$P(\mathbf{x}_i, \mathbf{z}_i \mid \theta) = \sum_{j=1}^{K} z_{ij}\tau_j \frac{1}{(2\pi)^{d/2}\sqrt{|\boldsymbol{\Sigma}_j|}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_i)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j)\right). \qquad (9.10)$$

Здесь θ – конгломерат все параметров τ, μ_1, ..., μ_K и $\boldsymbol{\Sigma}_1$, ..., $\boldsymbol{\Sigma}_K$. Интерпретация в качестве порождающей модели следующая: сначала случайным образом выбираем гауссиану, применяя априорное распределение τ, а затем используем выбранную гауссиану с помощью индикаторных переменных z_{ij}.

Для применения EM-алгоритма выписываем Q-функцию:

$$\begin{aligned}
Q(\theta \mid \theta^t) &= \mathbb{E}[\ln P(\mathbf{X} \cup \mathbf{Z} \mid \theta) \mid \mathbf{X}, \theta^t] = \\
&= \mathbb{E}\left[\ln \prod_{i=1}^{n} P(\mathbf{x}_i \cup \mathbf{z}_i \mid \theta) \middle| \mathbf{X}, \theta^t\right] = \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \ln P(\mathbf{x}_i \cup \mathbf{z}_i \mid \theta) \middle| \mathbf{X}, \theta^t\right] = \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \ln \sum_{j=1}^{K} z_{ij}\tau_j \frac{1}{(2\pi)^{d/2}\sqrt{|\boldsymbol{\Sigma}_j|}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j)\right) \middle| \mathbf{X}, \theta^t\right] = \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij} \ln\left(\tau_j \frac{1}{(2\pi)^{d/2}\sqrt{|\boldsymbol{\Sigma}_j|}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j)\right)\right) \middle| \mathbf{X}, \theta^t\right] = \quad (*) \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\left(\ln\tau_j - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}_i - \mu_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j)\right) \middle| \mathbf{X}, \theta^t\right] = \\
&= \sum_{i=1}^{n}\sum_{j=1}^{K} \mathbb{E}[z_{ij} \mid \mathbf{X}, \theta^t]\left(\ln\tau_j - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}_i - \mu_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j)\right).
\end{aligned} \qquad (9.11)$$

Шаг, помеченный звездочкой (*), возможен, потому что для любого i среди чисел z_{ij} только одно отлично от 0, следовательно, мы можем вынести индикаторные переменные из-под знака логарифма. В последней строке Q-функция представлена в желаемой форме, включающей, с одной стороны, математические ожидания скрытых переменных при условии наблюдаемых данных $\mathbf{X}$ и ранее

оцененных параметров θ^t, а с другой – зависящими от θ выражениями, которые позволяют найти θ^{t+1} путем максимизации.

Шаг Expectation EM-алгоритма сводится, таким образом, к вычислению ожидаемых значений индикаторных переменных $\mathbb{E}[z_{ij}|\mathbf{X},\theta^t]$. Отметим, что математические ожидания булевых переменных принимают значения во всем отрезке [0, 1] с тем ограничением, что $\Sigma_{j=1}^K z_{ij} = 1$ для всех i. По существу, жесткое назначение кластеров в алгоритме K средних заменяется мягким назначением – это один из видов обобщения алгоритма K средних с помощью гауссовых смесовых моделей. Теперь предположим, что $K = 2$ и что мы ожидаем, что кластеры имеют одинаковые размеры и одинаковые ковариации. Если заданная точка $\mathbf{x}_i$ равноудалена от средних точек обоих кластеров (а точнее, наших оценок этих точек), то очевидно, что $\mathbb{E}[z_{i1}|\mathbf{X},\theta^t] = \mathbb{E}[z_{i2}|\mathbf{X},\theta^t] = 1/2$. В общем случае эти математические ожидания распределяются пропорционально массе вероятности, назначенной этой точке каждой гауссианой:

$$\mathbb{E}[z_{ij}\,|\,\mathbf{X},\theta^t] = \frac{\tau_j^t f(\mathbf{x}_i\,|\,\mu_j^t,\boldsymbol{\Sigma}_j^t)}{\Sigma_{k=1}^K \tau_k^t f(\mathbf{x}_i\,|\,\mu_k^t,\boldsymbol{\Sigma}_k^t)}, \tag{9.12}$$

где $f(\mathbf{x}|\mu, \boldsymbol{\Sigma})$ – многомерная гауссова функция плотности.

На шаге Maximisation мы оптимизируем параметры в уравнении (9.11). Отметим, что не существует никакой связи между членами, содержащими τ_j, и членами, содержащими другие параметры, поэтому априорное распределение τ можно оптимизировать отдельно:

$$\tau^{t+1} = \arg\max_{\tau} \sum_{i=1}^{n}\sum_{j=1}^{K} \mathbb{E}[z_{ij}\,|\,\mathbf{X},\theta^t]\ln\tau_j =$$

$$= \arg\max_{\tau} \sum_{j=1}^{K} E_j \ln\tau_j \quad \text{с ограничением } \sum_{j=1}^{K}\tau_j = 1,$$

где E_j обозначает сумму $\Sigma_{i=1}^n \mathbb{E}[z_{ij}|\mathbf{X},\theta^t]$ – полное членство j-го кластера; отметим, что $\Sigma_{j=1}^K E_j = n$. Для простоты предположим, что $K = 2$, так что $\tau_2 = 1 - \tau_1$, тогда

$$\tau_1^{t+1} = \arg\max_{\tau_1} E_1 \ln\tau_1 + E_2 \ln(1-\tau_1).$$

Приравнивая к нулю частную производную по τ_1 и решая получившееся уравнение относительно τ_1, легко проверить, что $\tau_1^{t+1} = E_1/(E_1 + E_2) = E_1/n$ и, следовательно, $\tau_2^{t+1} = E_2/n$. В общем случае K кластеров имеем аналогично:

$$\tau_j^{t+1} = \frac{E_j}{\Sigma_{k=1}^K E_k} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_{ij}\,|\,\mathbf{X},\theta^t]. \tag{9.13}$$

Средние и ковариационные матрицы можно оптимизировать для каждого кластера по отдельности:

$$\mu_j^{t+1}, \boldsymbol{\Sigma}_j^{t+1} = \arg\max_{\mu_j, \boldsymbol{\Sigma}_j} \sum_{i=1}^{n} \mathbb{E}[z_{ij} \,|\, \mathbf{X}, \theta^t] \left(-\frac{1}{2} \ln |\, \boldsymbol{\Sigma}_j \,| - \frac{1}{2} (\mathbf{x}_i - \mu_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_i - \mu_j) \right) =$$

$$= \arg\max_{\mu_j, \boldsymbol{\Sigma}_j} \sum_{i=1}^{n} \mathbb{E}[z_{ij} \,|\, \mathbf{X}, \theta^t] \left(\frac{1}{2} \ln |\, \boldsymbol{\Sigma}_j \,| + \frac{1}{2} (\mathbf{x}_i - \mu_j)^{\mathrm{T}} \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_i - \mu_j) \right).$$

Отметим, что выражение в скобках – квадрат расстояния, а математические ожидания играют роль весов, соответствующих каждому объекту. Эта формула описывает обобщенный вариант задачи нахождения точки, которая ☞ *обращает в минимум сумму квадратов евклидовых расстояний* до множества точек (теорема 8.1 на стр. 249). И если решением той задачи являлось среднее арифметическое, то здесь мы просто берем *взвешенное* среднее по всем точкам:

$$\mu_j^{t+1} = \frac{1}{E_j} \sum_{i=1}^{n} \mathbb{E}[z_{ij} \,|\, \mathbf{X}, \theta^t] \, \mathbf{x}_i = \frac{\Sigma_{i=1}^{n} \mathbb{E}[z_{ij} \,|\, \mathbf{X}, \theta^t] \, \mathbf{x}_i}{\Sigma_{i=1}^{n} \mathbb{E}[z_{ij} \,|\, \mathbf{X}, \theta^t]}. \tag{9.14}$$

Аналогично ковариационная матрица вычисляется как взвешенное среднее ковариационных матриц по всем точкам, с учетом новой оценки среднего:

$$\boldsymbol{\Sigma}_j^{t+1} = \frac{1}{E_j} \sum_{i=1}^{n} \mathbb{E}[z_{ij} \,|\, \mathbf{X}, \theta^t] (\mathbf{x}_i - \mu_j^{t+1}) (\mathbf{x}_i - \mu_j^{t+1})^{\mathrm{T}} =$$

$$= \frac{\Sigma_{i=1}^{n} \mathbb{E}[z_{ij} \,|\, \mathbf{X}, \theta^t] (\mathbf{x}_i - \mu_j^{t+1}) (\mathbf{x}_i - \mu_j^{t+1})^{\mathrm{T}}}{\Sigma_{i=1}^{n} \mathbb{E}[z_{ij} \,|\, \mathbf{X}, \theta^t]}. \tag{9.15}$$

Уравнения (9.12)–(9.15) составляют полученное с помощью EM-алгоритма решение для обучения гауссовой смесовой модели по непомеченной выборке. Я представил его в самой общей форме, когда явно моделируются неравные размеры кластеров и различные ковариационные матрицы. Последнее важно, так как позволяет иметь кластеры различной формы – в отличие от алгоритма *K* средних, в котором предполагается, что все кластеры имеют одинаковую сферическую форму. Следовательно, границы между кластерами не будут линейными, как в случае кластеризации методом *K* средних. На рис. 9.8 демонстрируется сходимость EM-алгоритма на простом одномерном наборе данных, а также существование нескольких стационарных состояний.

В заключение повторим, что алгоритм Expectation-Maximisation – гибкий и эффективный способ решения задач с отсутствующими данными, имеющий твердые теоретические основания. Как мы видели на примере гауссовой смесовой модели, основным его ингредиентом является выражение для параметрической функции правдоподобия $P(X \cup Z|\theta)$, из которого с помощью Q-функции можно вывести уравнения обновления. Нелишним будет предостережение о том, что, за исключением простейших случаев, стационарных состояний может быть несколько. Поэтому, как и в случае алгоритма *K* средних, оптимизацию следует выполнять несколько раз с различными начальными значениями.

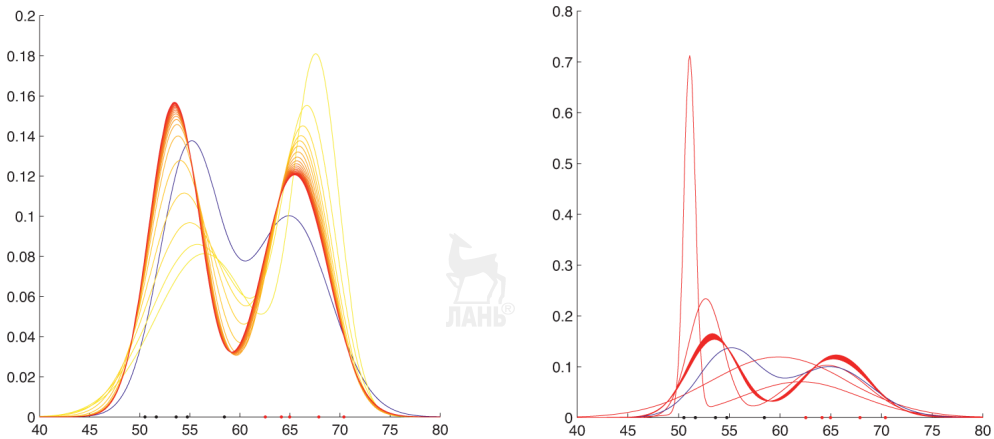

Рис. 9.8. (Слева) Синей линией показана истинная гауссова смесовая модель, из которой были выбраны 10 точек на оси x; цвет точек обозначает, из какой гауссианы они взяты: левой или правой. Остальные линии демонстрируют сходимость EM-алгоритма к стационарному состоянию после случайной инициализации. **(Справа)** На этом графике показаны четыре стационарных состояния для одного и того же набора данных. Было выполнено 20 итераций EM-алгоритма; большая толщина одной из линий говорит о том, что в этой конфигурации время сходимости оказалось дольше

9.5 Модели на основе сжатия

Мы закончим эту главу кратким обсуждением подхода к машинному обучению, который одновременно тесно связан и разительно отличается от вероятностного.

$$y_{\text{MAP}} = \arg\max_y P(X = x \,|\, Y = y)P(Y = y).$$

Взяв логарифмы со знаком минус, мы можем преобразовать эту задачу к эквивалентной задаче минимизации:

$$y_{\text{MAP}} = \arg\min_y -\log P(X = x \,|\, Y = y) - \log P(Y = y). \tag{9.16}$$

Это следует из того, что для любых двух вероятностей $0 < p < p' < 1$ имеем $\infty > -\log p > -\log p' > 0$. Если вероятность события равна p, то логарифм p со знаком минус дает количественное выражение *объема информации*, содержащейся в сообщении о том, что событие произошло. Это интуитивно понятно, потому что чем менее ожидаемо событие, тем больше информации содержит объявление о нем. Единица информации зависит от основания логарифма: принято брать логарифмы по основанию 2, и в таком случае объем информации измеряется в битах. Например, если вы один раз подбросите правильную монету и скажете мне, что выпала решка, то в этом сообщении будет содержаться $-\log_2 1/2 =$

= 1 бит информации; если вы один раз бросите правильную кость и скажете, что выпала шестерка, то объем информации в сообщении будет равен $-\log_2 1/6 = 2.6$ бита. Уравнение (9.16) говорит, что решающее правило MAP выбирает наименее неожиданный, или самый ожидаемый, класс объекта x при условии заданных априорных распределений и правдоподобий. Мы пишем $IC(X|Y) = -\log_2 P(X|Y)$ и $IC(Y) = -\log_2 P(Y)$[1].

Пример 9.7 (классификация на основе информации). В табл. 9.2 воспроизведена левая часть табл. 1.3 на стр. 41 вместе с соответствующими количественными показателями объема информации. Если Y распределена равномерно, то $IC(Y = \text{спам}) = 1$ бит и $IC(Y = \text{неспам}) = 1$ бит. Отсюда следует, что

$$\arg\min_y (IC(\text{виагра} = 1|Y = y) + IC(Y = y)) = \text{спам};$$
$$\arg\min_y (IC(\text{виагра} = 0|Y = y) + IC(Y = y)) = \text{неспам}.$$

Если неспам в четыре раза вероятнее спама, то $IC(Y = \text{спам}) = 2.32$ бита, $IC(Y = \text{неспам}) = 0.32$ бита и $\text{argmin}_y (IC(\text{виагра} = 1|Y = y) + IC(Y = y)) = \text{неспам}$.

Y	$P(\text{виагра} = 1\|Y)$	$IC(\text{виагра} = 1\|Y)$	$P(\text{виагра} = 0\|Y)$	$IC(\text{виагра} = 0\|Y)$
Спам	0.40	**1.32 бита**	0.60	**0.74 бита**
Неспам	0.12	**3.06 бита**	0.88	**0.18 бита**

Таблица 9.2. Примеры маргинальных правдоподобий

Очевидно, что при равномерном распределении k исходов каждый исход содержит один и тот же объем информации $-\log_2 1/k = \log_2 k$. Если распределение неравномерно, то объем информации будет различаться, поэтому имеет смысл вычислить средний объем информации, или *энтропию* $\sum_{i=1}^{k} -p_i \log_2 p_i$. Мы уже встречались с энтропией в разделе 5.1, только тогда называли ее ☞ *мерой нечистоты*.

До сих пор в том, что я сказал, нового было разве что наличие взаимно однозначного соответствия между вероятностью и объемом информации. А настоящим двигателем обучения на основе сжатия является фундаментальный результат из теории информации, доказанный в 1948 году Клодом Шенноном. Результат Шеннона гласит – без излишнего формализма, – что невозможно передавать информацию со скоростью, большей энтропии, но с помощью хитроумных двоичных кодов можно достичь скорости, сколь угодно близкой к оптимальной. К числу хорошо известных относятся коды Шеннона-Фано и Хаффмана, с которыми стоит познакомиться, потому что в них используется простая древовидная структура для построения кода на основе эмпирических вероятностей. Существуют и более эффективные коды, например арифметическое кодирование, в которых несколько сообщений объединяются в одно кодовое слово.

[1] Здесь IC – сокращение от Information Content (объем информации). – *Прим. перев.*

Предполагая наличие почти оптимального кода, мы можем сменить точку зрения и использовать объем информации – или, как его чаще называют, «длину описания» – вместо вероятности. Упрощенный вариант принципа минимальной длины описания (МДО) формулируется следующим образом.

Определение 9.1 (принцип минимальной длины описания). *Обозначим $L(m)$ длину в битах описания модели m, а $L(D|m)$ – длину в битах описания данных D при условии модели m. Согласно принципу минимальной длины описания, предпочтительной является модель, минимизирующая сумму длин описания модели и данных при условии модели:*

$$m_{\text{MDL}} = \arg\min_{m \in M}(L(m) + L(D|m)). \tag{9.17}$$

В контексте прогностического обучения «описание данных при условии модели» относится к той информации – помимо самой модели и значений признаков данных, – которая необходима для вывода целевых меток. Если модель на 100% точна, то никакой дополнительной информации не нужно, так что этот член, по существу, служит для количественного выражения степени некорректности модели. Например, в случае двух равномерно распределенных классов нам нужен один бит для каждого объекта, неправильно классифицированного моделью. Член $L(m)$ является количественным выражением сложности модели. Например, если мы аппроксимируем данные полиномом, то должны закодировать степень полинома и его корни с некоторой разрешающей способностью. Таким образом, принцип МДО устанавливает компромисс между верностью и сложностью модели: член, описывающий сложность, позволяет избежать переобучения – как ☞ *регуляризующий член* в гребневой регрессии (раздел 7.1) и ☞ *ослабляющая переменная* в методе опорных векторов с мягким зазором (раздел 7.3).

Какое кодирование использовать для определения сложности модели $L(m)$ – вопрос, часто неочевидный и в какой-то мере субъективный. Тут можно провести аналогию с байесовским подходом, когда мы должны определить априорное распределение моделей. Подход на основе МДО предлагает конкретный способ определения априорного распределения моделей с помощью кодов.

9.6 Вероятностные модели: итоги и литература для дальнейшего чтения

В этой главе мы рассмотрели ряд моделей машинного обучения, в основе которых лежит идея моделирования признаков и целевых переменных случайными величинами, что дает возможность явно представлять и манипулировать уровнем доверия к этим величинам. Такие модели обычно являются прогностическими в том смысле, что их результатом является условное распределение вероятности $P(Y|X)$, с помощью которого Y можно предсказать, зная X. Порождающие мо-

дели оценивают совместное распределение $P(Y, X)$ – часто с помощью функции правдоподобия $P(X|Y)$ и априорного распределения $P(Y)$, – из которого можно получить апостериорное распределение $P(Y|X)$. Напротив, в условных моделях обученное апостериорное распределение $P(Y|X)$ получается непосредственно, без затрат ресурсов на обучение $P(X)$. Для «байесовского» подхода к машинному обучению характерно стремление получить полное апостериорное распределение, когда это осуществимо, а не просто найти максимизирующее значение.

- В разделе 9.1 мы видели, что нормальное, или гауссово, распределение поддерживает многие полезные геометрические представления, прежде всего потому, что логарифм со знаком минус гауссова правдоподобия можно интерпретировать как квадрат расстояния. Одинаковые ковариационные матрицы классов приводят к прямолинейным решающим границам, а это означает, что модели, дающие такие линейные границы, в том числе линейные классификаторы, линейную регрессию и кластеризацию методом K средних, можно интерпретировать с вероятностной точки зрения, что делает встроенные в них предположения явными. Мы рассмотрели два иллюстративных примера: (*i*) базовый линейный классификатор является оптимальным по Байесу в случае некоррелированных гауссовых признаков с единичной дисперсией и (*ii*) регрессия по методу наименьших квадратов оптимальна для линейных функций, загрязненных гауссовым шумом.
- Раздел 9.2 был посвящен различным вариантам наивного байесовского классификатора, в котором делается упрощающее предположение о независимости признаков внутри каждого класса. Обзор и история вопроса приведены в работе Lewis (1998). Эта модель широко используется для информационного поиска и классификации текстов, поскольку она часто дает хорошее ранжирование, пусть даже порождаемые ей оценки вероятностей далеки от совершенства. Хотя обычно в модели, именуемой наивной, признаки рассматриваются как категориальные, или случайные, величины с распределением Бернулли, варианты, в которых применяется мультиномиальное распределение, как правило, лучше моделируют количество вхождений слов в документ (McCallum, Nigam, 1998). Учесть вещественные признаки можно двумя способами: моделируя их как нормально распределенные в пределах каждого класса или с помощью непараметрического оценивания плотности, – в работе John, Langley (1995) утверждается, что последний способ дает лучшие эмпирические результаты. В работе Webb, Boughton, Wang (2005) обсуждается, как можно ослабить требование независимости, лежащее в основе наивной байесовской классификации. Сглаживание вероятностей с помощью m-оценки было впервые предложено в работе Cestnik (1990).
- Хотя это и звучит парадоксально, я не считаю, что в наивном байесовском классификаторе есть что-то «байесовское». Да, это порождающая вероятностная модель для оценивания апостериорного распределения $P(Y|X)$ по совместному распределению $P(Y, X)$, но на практике апостериорное

распределение очень плохо откалибровано из-за нереалистичных предположений о независимости. В результате анализа, проведенного в работе Domingos, Pazzani (1997), установлено, что наивный байесовский классификатор так часто оказывается успешным вследствие высокого качества $\text{argmax}_Y P(Y|X)$, а не апостериорного распределения как такового. Более того, даже без использования правила Байеса, при определении доставляющего максимум значения Y, можно обойтись, поскольку оно служит лишь для преобразования неоткалиброванных правдоподобий в неоткалиброванные апостериорные вероятности. Поэтому я рекомендую использовать наивные байесовские правдоподобия как оценки с неизвестной шкалой, для которых порог принятия решения следует откалибровать посредством анализа кривой PXП, как обсуждалось в нескольких местах выше.

- В разделе 9.3 мы рассмотрели широко распространенную модель логистической регрессии. Основная идея – объединить линейную решающую границу с логистической калибровкой, но обучить ее в дискриминантной манере путем оптимизации условного правдоподобия. Таким образом, вместо моделирования классов в виде облаков точек и вывода из них решающей границы логистическая регрессия сосредоточивается на областях перекрытия классов. Это пример более широкого класса обобщенных линейных моделей (Nelder, Wedderburn, 1972). В работе Jebara (2004) обсуждаются преимущества дискриминантного обучения, по сравнению с порождающими моделями. Дискриминантное обучение применимо также к последовательным данным в форме условных случайных полей (Lafferty et al., 2001).
- В разделе 9.4 был в общем виде представлен алгоритм Expectation-Maximisation как способ обучения моделей со скрытыми переменными. Эта общая форма EM-алгоритма предложена в работе Dempster, Laird, Rubin (1977), основанной на более ранних работах. Мы видели, как ее можно применить к гауссовым смесовым моделям для получения более общего варианта прогностической кластеризации методом K средних, который умеет также оценивать формы и размеры кластеров. Однако при этом увеличивается число параметров модели, а значит, и риск застрять в неоптимальном стационарном состоянии. Работы Little, Rubin, 1987 – стандартный источник по работе в условиях отсутствия данных.
- Наконец, в разделе 9.5 мы вкратце обсудили некоторые идеи, относящиеся ко взгляду на обучение как на сжатие. Связь с вероятностным моделированием заключается в том, что в обоих случаях цель состоит в моделировании и использовании неслучайных аспектов данных. В упрощенной форме принцип минимальной длины описания можно вывести из правила Байеса, взяв логарифм со знаком минус; он утверждает, что следует предпочесть модель, которая минимизирует сумму длин описания модели и данных при условии модели. Первый член количественно выражает сложность модели, а второй – ее верность (поскольку в явном кодировании нуждаются лишь ошибки модели). Достоинство принципа МДО заклю-

чается в том, что схемы кодирования зачастую более осязаемы, их проще определить, чем априорные распределения вероятностей. Однако подойдет не любое кодирование: как и в случае вероятностей, для этих схем необходимо обоснование в моделируемой предметной области. Основополагающими в этой области были работы Solomonoff (1964 a, b); Wallace, Boulton (1968); Rissanen (1978) и другие. Отличным введением и обзором может послужить работа Grunwald (2007).