Курс: Байесовские методы машинного обучения, 2011

EM-алгоритм и метод релевантных векторов для классификации

Дата: 2 ноября 2011

Метод оптимизации Ньютона

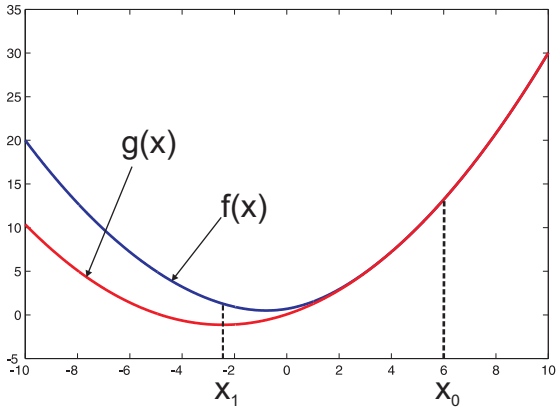

Рис. 1: Иллюстрация одной итерации метода Ньютона для минимизации функции $f(x) = \log(1 + \exp(x)) + x^2/5$ путем аппроксимации с помощью квадратичной функции g(x).

Рассмотрим задачу безусловной оптимизации гладкой функции

$$f(\boldsymbol{x}) \to \min_{\boldsymbol{x}}, \quad \boldsymbol{x} \in \mathbb{R}^d.$$

Идея метода оптимизации Ньютона заключается в итерационном приближении функции $f(\boldsymbol{x})$ с помощью квадратичной функции $g(\boldsymbol{x})$, минимум которой может быть найден аналитически. Приближение функции с помощью квадратичной функции путем приравнивания нулевой, первой и второй производных в точке $\boldsymbol{x}_0$ эквивалентно разложению в ряд Тейлора функции f с отбрасыванием всех членов третьего и более высокого порядков:

$$f(\boldsymbol{x}) \simeq g(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{x}_0))^T(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T \nabla\nabla f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0).$$

Здесь через ∇f и $\nabla\nabla f$ обозначены градиент и гессиан функции f соответственно. Минимум функции g может быть легко найден путем приравнивания градиента g к нулю:

$$\nabla g(\boldsymbol{x}_*) = \nabla f(\boldsymbol{x}_0) + \nabla\nabla f(\boldsymbol{x}_0)(\boldsymbol{x}_* - \boldsymbol{x}_0) = \mathbf{0} \;\Rightarrow\; \boldsymbol{x}_* = \boldsymbol{x}_0 - (\nabla\nabla f(\boldsymbol{x}_0))^{-1}\nabla f(\boldsymbol{x}_0).$$

Таким образом, метод оптимизации Ньютона представляет собой следующий итерационный процесс (см. рис. 1):

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - (\nabla\nabla f(\boldsymbol{x}_t))^{-1}\nabla f(\boldsymbol{x}_t). \tag{1}$$

Метод Ньютона имеет на порядок более высокую скорость сходимости по сравнению с методом градиентного спуска, но при этом гораздо меньшую область применимости. Сходимость метода Ньютона доказана для выпуклых функций с ограниченным модулем производной (в более общем случае для Липшиц-непрерывных функций с обратимым гессианом). К сожалению, на практике такие функции встречаются нечасто. Тем не менее, во многих случаях метод Ньютона можно применять при условии выбора хорошего начального приближения $\boldsymbol{x}_0$, в окрестности которого функция удовлетворяет условиям выпуклости и Липшиц-непрерывности.

На каждом шаге метода Ньютона (1) требуется вычислять и обращать гессиан оптимизируемой функции. В пространствах большой размерности поиск гессиана может быть связан со значительными вычислительными сложностями. Кроме того, обращение гессиана, близкого к вырожденному,

сопряжено с численными неустойчивостями. В этих случаях используются т.н. квази-ньютоновские методы (например, алгоритм Левенберга-Марквардта, L-BFGS и другие), которые используют различные приближения гессиана и вводят поправки на вырожденный случай.

EM-алгоритм в общем виде

Пусть имеется вероятностная модель, задаваемая совместным распределением $p(X,T|\Theta)$. Здесь X – набор наблюдаемых переменных, T – набор ненаблюдаемых переменных и Θ – набор параметров модели. Рассмотрим задачу обучения модели (поиск параметров Θ по выборке X) с помощью метода максимального правдоподобия:

$$\log p(X|\Theta) = \log \int p(X,T|\Theta)dT \to \max_{\Theta}. \tag{2}$$

Рассмотрим произвольное вероятностное распределение $q(T)$. Тогда справедлива следующая цепочка равенств:

$$\log p(X|\Theta) = \int q(T)\log p(X|\Theta)dT = \int q(T)\log\frac{p(X,T|\Theta)}{p(T|X,\Theta)}dT = \int q(T)\log\left[\frac{p(X,T|\Theta)}{q(T)}\frac{q(T)}{p(T|X,\Theta)}\right]dT =$$
$$= \underbrace{\int q(T)\log p(X,T|\Theta)dT - \int q(T)\log q(T)dT}_{\mathcal{L}(q)} \underbrace{- \int q(T)\log\frac{p(T|X,\Theta)}{q(T)}dT}_{\mathrm{KL}(q||p(T|X,\Theta))}.$$

Дивергенция $\mathrm{KL}(q||p(T|X,\Theta)) \geq 0$, следовательно

$$\log p(X|\Theta) \geq \mathcal{L}(q). \tag{3}$$

При этом нижняя граница (3) становится точной, если $q(T) = p(T|X,\Theta)$, т.к. дивергенция $\mathrm{KL}(q||p(T|X,\Theta)) = 0$ для тождественных распределений.

EM-алгоритм для решения задачи (2) представляет собой аналог метода Ньютона для оптимизации произвольной функции, где вместо квадратичного приближения в текущей точке используется нижняя оценка (3). Пусть фиксировано некоторое значение параметров Θ_{old}. Сначала (на E-шаге) находится распределение $q(T)$ как распределение значений скрытых переменных при данных параметрах:

$$q(T) = p(T|X,\Theta_{old}) = \frac{p(X,T|\Theta_{old})}{\int p(X,T|\Theta_{old})dT}. \tag{4}$$

При таком выборе $q(T)$ нижняя оценка (3) является точной при $\Theta = \Theta_{old}$. Затем (на M-шаге) новые значения параметров Θ находятся путем максимизации нижней границы $\mathcal{L}(q)$, что эквивалентно решению задачи

$$\mathbb{E}_{T|X,\Theta_{old}} \log p(X,T|\Theta) \to \max_{\Theta}, \tag{5}$$

так как энтропия $-\mathbb{E}_q \log q(T)$ распределения $q(T)$ не зависит от Θ. Шаги E и M повторяются в цикле до сходимости. Очевидно, что в процессе EM-итераций нижняя оценка (3), а также значение правдоподобия $p(X|\Theta)$ не убывают. Монотонное неубывание правдоподобия $p(X|\Theta)$ и его ограниченность сверху гарантируют общую сходимость EM-итераций.

Итерационный процесс в EM-алгоритме проиллюстрирован на рис. 2. По аналогии со многими итерационными методами оптимизации, EM-алгоритм позволяет находить только локальный максимум правдоподобия.

Заметим, что во многих практических случаях решение задачи (5) намного проще, чем решение задачи (2). В частности, в рассматриваемой ниже задаче разделения гауссовской смеси задача оптимизации на M шаге может быть решена аналитически.

Вычисление значения функции правдоподобия $p(X|\Theta)$ в фиксированной точке Θ требует интегрирования по пространству T и в ряде случаев может представлять собой вычислительно трудоемкую задачу. Заметим, что эта величина правдоподобия необходима также для вычисления апостериорного распределения $p(T|X,\Theta_{old})$ на E шаге. Однако, распределение $p(T|X,\Theta_{old})$ используется затем только для вычисления математического ожидания логарифма полного правдоподобия на M шаге. Как правило, здесь не требуется знать все апостериорное распределение целиком, а

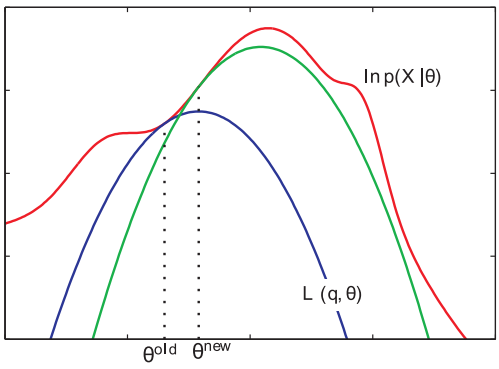

Рис. 2: Иллюстрация итерационного процесса в EM-алгоритме. Нижняя оценка (3) обозначена через $L(q, \theta)$.

достаточно знать лишь несколько статистик этого распределения (например, только мат.ожидания отдельных компонент $\mathbb{E}_{T|X,\Theta_{old}} t_n$ и парные ковариации $\mathbb{E}_{T|X,\Theta_{old}} t_n t_k$). Поэтому EM-алгоритм может быть вычислительно эффективен даже в тех случаях, когда вычисление значения правдоподобия $p(X|\Theta)$ в одной точке затруднено.

EM-алгоритм можно рассматривать как покоординатный подъем для максимизации нижней границы (3). Сначала при фиксированных параметрах Θ_{old} нижняя граница максимизируется по распределению $q(T)$ (эта задача имеет аналитическое решение $q(T) = p(T|X, \Theta_{old})$). Затем при фиксированном $q(T)$ нижняя граница максимизируется по параметрам Θ. Если апостериорное распределение $p(T|X, \Theta_{old})$ не поддается вычислению, то можно ограничить семейство распределений $q(T)$ параметрическим или факторизованным семейством и решать задачу максимизации нижней границы в рамках выбранного семейства распределений. В этом случае будет обеспечиваться монотонный рост нижней границы правдоподобия, но, вообще говоря, не самого значения правдоподобия $p(X|\Theta)$. Такой подход будет обсуждаться в рамках вариационного подхода к решению задачи приближенного байесовского вывода. В случае невозможности точного решения задачи на M шаге можно ограничиться лишь движением в сторону возрастания нижней границы.

EM-алгоритм можно применять также для решения задачи обучения вероятностной модели со скрытыми переменными $p(X, T|\Theta)$ с помощью максимизации апостериорного распределения:

$$p(\Theta|X) \to \max_{\Theta} \;\Leftrightarrow\; \log p(X|\Theta) + \log p(\Theta) \to \max_{\Theta}.$$

Здесь $p(\Theta)$ — априорное распределение на параметры модели. В этом случае нижней оценкой для оптимизируемого функционала является следующая величина:

$$\log p(X|\Theta) + \log p(\Theta) \geq \mathcal{L}(q) + \log p(\Theta).$$

Итерационная оптимизация данной нижней границы по распределению $q(T)$ и по параметрам Θ приводит к E шагу (4) и M шагу

$$\mathbb{E}_{T|X,\Theta_{old}} \log p(X, T|\Theta) + \log p(\Theta) \to \max_{\Theta}.$$

EM-алгоритм для разделения гауссовской смеси

Рассмотрим вероятностную модель смеси нормальных распределений:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k),\; \sum_{k=1}^{K} w_k = 1,\; w_k \geq 0. \tag{6}$$

Модель смеси распределений (не обязательно нормальных) можно рассматривать как модель со скрытой переменной t, которая обозначает номер компоненты смеси:

$$p(t = k) = w_k, \tag{7}$$
$$p(\boldsymbol{x}|t = k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k). \tag{8}$$

Легко показать, что маргинальное распределение $p(\boldsymbol{x}) = \sum_k p(\boldsymbol{x}|t=k)p(t=k)$ в этой модели совпадает с распределением (6). В этом смысле модели (7)-(8) и (6) эквивалентны.

Интерпретация вероятностной модели смеси распределений как модели со скрытой переменной позволяет генерировать выборку из модели смеси следующим образом. Сначала с вероятностями, равными $\boldsymbol{w}$, генерируется номер компоненты смеси, из которой затем генерируется точка $\boldsymbol{x}$.

Для аппроксимации выборки $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ с помощью модели смеси из K гауссиан можно воспользоваться методом максимального правдоподобия:

$$\begin{aligned} &p(X|\boldsymbol{w}, \{\boldsymbol{\mu}_k\}, \{\Sigma_k\}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{w}, \{\boldsymbol{\mu}_k\}, \{\Sigma_k\}) = \prod_{n=1}^{N}\left(\sum_k w_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)\right) \to \max_{\boldsymbol{w}, \{\boldsymbol{\mu}_k\}, \{\Sigma_k\}}, \\ &\sum_k w_k = 1,\ w_k \geq 0, \\ &\Sigma_k = \Sigma_k^T,\ \Sigma_k \succ 0. \end{aligned} \tag{9}$$

Данная задача условной оптимизации может быть эффективно решена с помощью EM-алгоритма, описанного выше. Заметим, что количество компонент смеси K не может быть найдено аналогичным образом с помощью максимизации правдоподобия, т.к. значение правдоподобия данных тем выше, чем больше компонент K используется. Для поиска оптимального значения K можно воспользоваться скользящим контролем, где критерием качества аппроксимации тестовых данных является значение правдоподобия.

При использовании EM-алгоритма для решения задачи (9) вероятностная модель смеси распределений интерпретируется как вероятностная модель со скрытыми переменными. Вычислим значение мат.ожидания логарифма полного правдоподобия, необходимого для решения задачи оптимизации на M шаге:

$$\mathbb{E}_q \log p(X, T|\boldsymbol{w}, \{\boldsymbol{\mu}_k\}, \{\Sigma_k\}) = \sum_{n=1}^{N}\sum_{k=1}^{K} q(t_n = k)(\log w_k + \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)). \tag{10}$$

Заметим, что это выражение зависит только от вероятностей отдельных скрытых переменных $q(t_n = k)$. Нетрудно показать, что эти величины можно найти следующим образом:

$$\gamma_{nk} \triangleq q(t_n = k) = p(t_n = k|\boldsymbol{x}_n, \boldsymbol{w}^{old}, \{\boldsymbol{\mu}_k^{old}\}, \{\Sigma_k^{old}\}) = \frac{w_k^{old}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{old}, \Sigma_k^{old})}{\sum_{j=1}^{K} w_j^{old}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j^{old}, \Sigma_j^{old})}. \tag{11}$$

Также нетрудно показать, что задача максимизации критерия (10) при ограничениях $\sum_{k=1}^{K} w_k = 1,\ w_k \geq 0$ может быть решена аналитически:

$$w_k = \frac{1}{N}\sum_{n=1}^{N} \gamma_{nk}, \tag{12}$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk}\boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma_{nk}}, \tag{13}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma_{nk}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \gamma_{nk}}. \tag{14}$$

Заметим, что решение для Σ_k (14) удовлетворяет условию симметричности и положительной определенности. Кроме того, формулы (13),(14) соответствуют оценкам максимального правдоподобия для многомерного нормального распределения, в которых каждый объект $\boldsymbol{x}_n$ берется с весом γ_{nk}.

Таким образом, EM-алгоритм для смеси нормальных распределений заключается в итерационном применении формул (11) и (12)-(14). Этот процесс имеет простую интерпретацию. Величина γ_{nk} показывает степень соответствия между объектом $\boldsymbol{x}_n$ и компонентой k (определяет вес объекта $\boldsymbol{x}_n$ для компоненты k). Эти веса затем используются на M шаге для вычисления новых значений параметров компонент. Иллюстрация применения EM-алгоритма для разделения нормальной смеси с двумя компонентами показана на рис. 3.

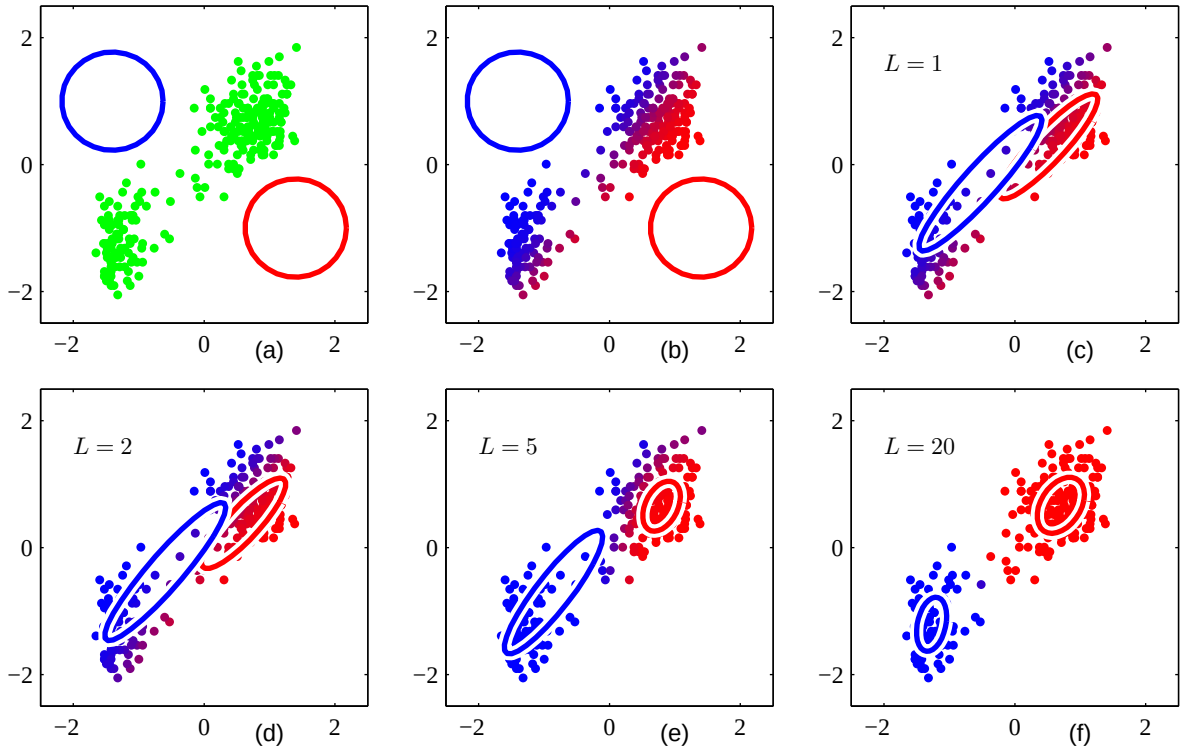

Рис. 3: Иллюстрация применения EM-алгоритма для разделения смеси нормальных распределений с двумя компонентами. На рис. a показана исходная выборка и начальное приближение для двух компонент. На рис. b показан результат E шага. При этом цвета объектов соответствуют значениям γ_{nk}. На рис. c-f показаны результаты вычислений после 1, 2, 5 и 20 итераций.

Одним из применений смеси нормальных распределений является решение задачи кластеризации на K кластеров. В этом случае номер кластера для объекта $\boldsymbol{x}_n$ определяется величиной

$$k_n = \arg\max_k \gamma_{nk}.$$

Такая схема кластеризации является вероятностным обобщением известного метода кластеризации K-средних.

В заключение этого раздела заметим, что восстановление плотности по данным (в частности, смеси нормальных распределений) является простейшим способом решения задачи идентификации. Для этого сначала для всех объектов обучающей выборки, обладающих заданным свойством, восстанавливается плотность распределения. Затем, для нового объекта $\boldsymbol{x}$ решение о наличии у него заданного свойства принимается, если значение восстановленной плотности $p(\boldsymbol{x})$ выше определенного порога.

Логистическая и мультиномиальная регрессия

Рассмотрим задачу классификации на два класса. Имеется обучающая выборка из N объектов $(T, X) = \{t_n, \boldsymbol{x}_n\}_{n=1}^N$, где $\boldsymbol{x}_n \in \mathbb{R}^d$ – вектор вещественных признаков для объекта n, а $t_n \in \{-1, 1\}$ – его метка класса. Задача заключается в том, чтобы определить метку класса t для нового объекта, представленного только своим вектором признаков $\boldsymbol{x}$.

Рассмотрим решение данной задачи в классе линейных решающих правил. Для этого введем функцию

$$f(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x}).$$

Здесь $w_j \in \mathbb{R}$ – некоторые веса, а $\phi_j : \mathbb{R}^d \to \mathbb{R}$ – фиксированные базисные функции, задаваемые пользователем. Примеры базисных функций:

- Исходные признаки: $\phi_j(\boldsymbol{x}) = x_j$;
- Полиномиальные базисные функции: $\phi_j(\boldsymbol{x}) = x_{j_1} x_{j_2} \dots x_{j_l},\ j_1, \dots, j_l \in \{1, \dots, d\},\ 1 \le l \le L$. L – степень полинома.
- Радиальные базисные функции: $\phi_j(\boldsymbol{x}) = \exp(-\gamma \|\boldsymbol{x} - \hat{\boldsymbol{x}}_j\|^2),\ \gamma > 0$. Здесь $\hat{\boldsymbol{x}}$ – некоторые фиксированные вектора в пространстве $\mathbb{R}^d$, например, объекты обучающей выборки или центры кластеров.

Решение о принадлежности объекта $\boldsymbol{x}$ к классу t принимается по знаку линейной функции f:

$$t(\boldsymbol{x}) = \begin{cases} +1, \text{ если } f(\boldsymbol{x}, \boldsymbol{w}) > 0, \\ -1, \text{ иначе.} \end{cases}$$

Метод классификации «Логистическая регрессия» представляет собой следующую вероятностную модель:

$$\begin{aligned} & p(T, \boldsymbol{w} | X, \alpha) = p(T | X, \boldsymbol{w}) p(\boldsymbol{w} | \alpha) = p(\boldsymbol{w} | \alpha) \prod_{n=1}^{N} p(t_n | \boldsymbol{x}_n, \boldsymbol{w}), \\ & p(t | \boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1 + \exp(-t f(\boldsymbol{x}, \boldsymbol{w}))}, \\ & p(\boldsymbol{w} | \alpha) = \mathcal{N}(\boldsymbol{w} | \boldsymbol{0}, \alpha^{-1} I). \end{aligned} \tag{15}$$

Здесь соответствие между меткой класса t и значением линейной функции f на объекте $\boldsymbol{x}$ вводится с помощью логистической функции. Гауссовское априорное распределение вводится для предотвращения переобучения, а параметр $\alpha \ge 0$ играет роль параметра регуляризации. Обучение в модели (15) (поиск весов $\boldsymbol{w}$) с помощью максимизации апостериорного распределения приводит к следующей задаче оптимизации:

$$\begin{aligned} p(\boldsymbol{w} | T, X, \alpha) \to \max_{\boldsymbol{w}} \ \Leftrightarrow \ \log p(T | X, \boldsymbol{w}) + \log p(\boldsymbol{w} | \alpha) \to \max_{\boldsymbol{w}} \ \Leftrightarrow \\ - \sum_{n=1}^{N} \log(1 + \exp(-t_n \sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x}_n))) - \frac{\alpha}{2} \|\boldsymbol{w}\|^2 \to \max_{\boldsymbol{w}}. \end{aligned} \tag{16}$$

Можно показать, что оптимизируемая функция в этой задаче является вогнутой по $\boldsymbol{w}$ (ее гессиан является неположительно определенным). Поэтому для решения задачи (16) можно воспользоваться методом оптимизации Ньютона (1). Параметр регуляризации α является структурным параметром, который может быть определен, например, с помощью скользящего контроля. Скалярный параметр в семействе радиальных или полиномиальных базисных функций также может быть определен с помощью скользящего контроля.

Пример применения логистической регрессии для задачи классификации с двумя признаками приведен на рис. 4a–d.

Рассмотрим задачу классификации на K классов. Имеется обучающая выборка $(T, X) = \{t_n, \boldsymbol{x}_n\}_{n=1}^{N}$, $\boldsymbol{x}_n \in \mathbb{R}^d$ – вектор признаков для объекта n, $t_n \in \{1, \dots, K\}$ – его метка класса. Задача состоит в определении метки класса t для нового объекта, представленного только своим вектором признаков $\boldsymbol{x}$.

Для решения данной задачи введем K линейных функций

$$f_k(\boldsymbol{x}, \boldsymbol{w}^k) = \sum_{j=1}^{M} w_j^k \phi_j(\boldsymbol{x}),\ k = 1, \dots, K.$$

Здесь $\boldsymbol{w}^k$ – веса k-ой линейной функции, а ϕ – общие базисные функции. Решение о принадлежности объекта $\boldsymbol{x}$ к классу t принимается по максимуму значений линейных функций $f_1, \dots, f_K$:

$$t(\boldsymbol{x}) = \arg\max_k f_k(\boldsymbol{x}, \boldsymbol{w}^k). \tag{17}$$

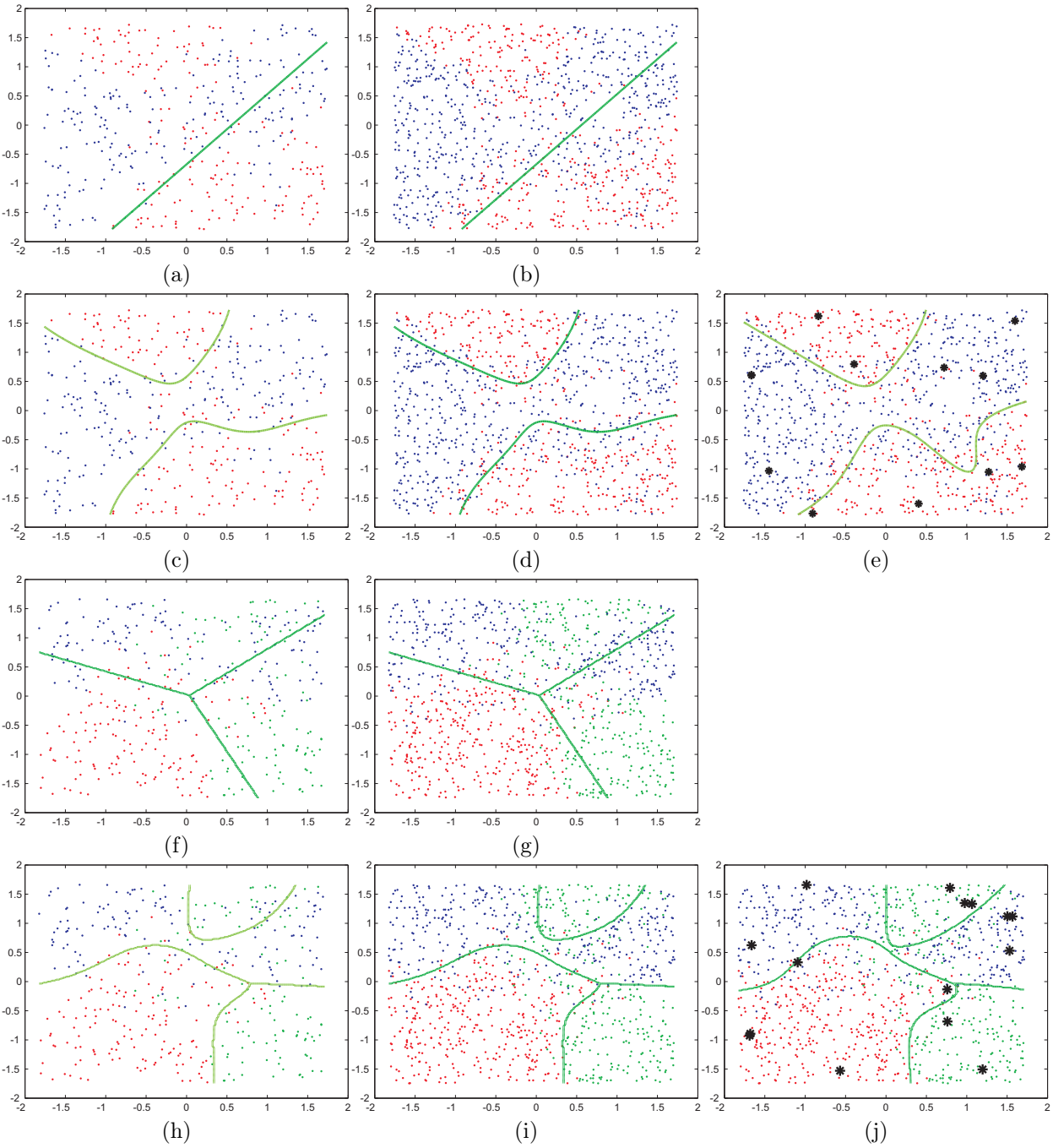

Рис. 4: Иллюстрация работы логистической и мультиномиальной регрессии, а также метода релевантных векторов. (a) – обучающая выборка с двумя классами, логистическая регрессия, исходные признаки, (b) – тестовая выборка с двумя классами, логистическая регрессия, исходные признаки, (c) – обучающая выборка с двумя классами, логистическая регрессия, радиальные базисные функции, (d) – тестовая выборка с двумя классами, логистическая регрессия, радиальные базисные функции, (e) – тестовая выборка с двумя классами, метод релевантных векторов, радиальные базисные функции, черными звездочками показаны релевантные центры радиальных базисных функций, (f) – обучающая выборка с тремя классами, мультиномиальная регрессия, исходные признаки, (g) – тестовая выборка с тремя классами, мультиномиальная регрессия, исходные признаки, (h) – обучающая выборка с тремя классами, мультиномиальная регрессия, радиальные базисные функции, (i) – тестовая выборка с тремя классами, мультиномиальная регрессия, радиальные базисные функции, (j) – тестовая выборка с тремя классами, многоклассовый метод релевантных векторов, радиальные базисные функции, черными звездочками показаны релевантные центры радиальных базисных функций.

По аналогии с логистической регрессией будем строить вероятностную модель, в которой функция соответствия между меткой класса t и набором значений линейных функций $f_1, \ldots, f_K$ выглядит как

$$p(t = k|\boldsymbol{x}, W) = p(t = k|f_1(\boldsymbol{x}, \boldsymbol{w}^1), \ldots, f_K(\boldsymbol{x}, \boldsymbol{w}^K)) = \frac{\exp(f_k(\boldsymbol{x}, \boldsymbol{w}^k))}{\sum_{j=1}^{K} \exp(f_j(\boldsymbol{x}, \boldsymbol{w}^j))}.$$

Здесь через $W \in \mathbb{R}^{K \times M}$ обозначена матрица весов всех линейных функций. Введенная функция называется мультиномиальной функцией, а основанная на ней вероятностная модель — «мультиномиальной регрессией»:

$$p(T, W|X, \alpha) = p(T|X, W)p(W|\alpha) = p(W|\alpha) \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n, W) = p(W|\alpha) \prod_{n=1}^{N} \frac{\exp(f_{t_n}(\boldsymbol{x}_n, \boldsymbol{w}^{t_n}))}{\sum_{j=1}^{K} \exp(f_j(\boldsymbol{x}_n, \boldsymbol{w}^j))},$$

$$p(W|\alpha) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{w}^k|\boldsymbol{0}, \alpha^{-1} I).$$

Здесь, как и в случае логистической регрессии, вводится гауссовское априорное распределение со скалярным параметром регуляризации α для предотвращения переобучения. Заметим, что модель мультиномиальной регрессии переходит в модель логистической регрессии для случая $K = 2$ и $f_2 = -f_1$.

Обучение в мультиномиальной регрессии (поиск весов W) по методу максимизации апостериорного распределения может быть осуществлено с помощью метода Ньютона решения задачи

$$\sum_{n=1}^{N} \left[f_{t_n}(\boldsymbol{x}_n, \boldsymbol{w}^{t_n}) - \log \left(\sum_{j=1}^{K} \exp(f_j(\boldsymbol{x}_n, \boldsymbol{w}^j)) \right) \right] - \frac{\alpha}{2} \sum_{j=1}^{K} \|\boldsymbol{w}^j\|^2 \to \max_W.$$

Параметр регуляризации α может быть настроен с помощью скользящего контроля. Пример применения мультиномиальной регрессии для задачи классификации с двумя признаками показан на рис. 4f–i.

Метод релевантных векторов для задачи классификации

Метод релевантных векторов (RVM) является надстройкой над вероятностной моделью логистической (для двух классов) и мультиномиальной (для многих классов) регрессии, позволяющей в процессе обучения отбирать информативные (релевантные) признаки (базисные функции).

Рассмотрим сначала метод RVM для случая двух классов. Вероятностная модель RVM выглядит следующим образом:

$$\begin{aligned} p(T, \boldsymbol{w}|X, \alpha) &= p(T|X, \boldsymbol{w})p(\boldsymbol{w}|\alpha) = p(\boldsymbol{w}|\alpha) \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n, \boldsymbol{w}), \\ p(t|\boldsymbol{x}, \boldsymbol{w}) &= \frac{1}{1 + \exp(-tf(\boldsymbol{x}, \boldsymbol{w}))}, \\ p(\boldsymbol{w}|\boldsymbol{\alpha}) &= \prod_{j=1}^{M} \mathcal{N}(w_j|0, \alpha_j^{-1}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \mathrm{diag}(\alpha)^{-1}). \end{aligned} \tag{18}$$

Как легко заметить, модель (18) полностью повторяет модель логистической регрессии (15), в которой матрица ковариации в гауссовском априорном распределении заменена на диагональную с различными элементами на диагонали. Таким образом, здесь вводится свой параметр регуляризации $\alpha_j \geq 0$ для каждого веса w_j. Обучение в модели (18) (поиск параметров $\boldsymbol{\alpha}$) производится с помощью метода максимального правдоподобия (обоснованности):

$$p(T|X, \boldsymbol{\alpha}) = \int p(T|X, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} = \int Q(\boldsymbol{w}, \boldsymbol{\alpha})d\boldsymbol{w} \to \max_{\boldsymbol{\alpha}}. \tag{19}$$

Здесь через $Q(\boldsymbol{w}, \boldsymbol{\alpha})$ обозначена подынтегральная функция. Интеграл в выражении для обоснованности (19) не вычисляется аналитически. Для приближенного оценивания данного интеграла в

RVM используется приближение Лапласа, при котором подынтегральная функция Q заменяется на свое ненормированное гауссовское приближение, интеграл от которого вычисляется аналитически:

$$p(T|X,\boldsymbol{\alpha}) \simeq \frac{Q(\boldsymbol{w}_{MP},\boldsymbol{\alpha})\sqrt{2\pi}^{M}}{\sqrt{\det \nabla^2_{\boldsymbol{w}} \log Q(\boldsymbol{w}_{MP},\boldsymbol{\alpha})}}, \quad \boldsymbol{w}_{MP} = \arg\max_{\boldsymbol{w}} Q(\boldsymbol{w},\boldsymbol{\alpha}). \tag{20}$$

Здесь через $\nabla^2_{\boldsymbol{w}}$ обозначен гессиан функции по $\boldsymbol{w}$. Считая $\boldsymbol{w}_{MP}$ фиксированным и приравнивая градиент полученного приближения обоснованности по $\boldsymbol{\alpha}$ к нулю, получаем следующую формулу пересчета для $\boldsymbol{\alpha}$:

$$\alpha_j^{new} = \frac{1-\alpha_j^{old}\Sigma_{jj}}{w^2_{MP,j}},\ \Sigma = (-\nabla^2_{\boldsymbol{w}} \log Q(\boldsymbol{w}_{MP},\alpha))^{-1}. \tag{21}$$

Таким образом, мы получаем итерационный алгоритм максимизации обоснованности (19). На каждой итерации сначала при текущем значении $\boldsymbol{\alpha}$ находится $\boldsymbol{w}_{MP}$ и строится приближение для обоснованности (20), а затем при текущем $\boldsymbol{w}_{MP}$ новое значение $\boldsymbol{\alpha}$ находится по формуле пересчета (21).

В процессе максимизации обоснованности (19) часть компонент вектора $\boldsymbol{\alpha}$ стремится к плюс бесконечности. В результате априорное распределение для соответствующих весов $\boldsymbol{w}$ переходит в дельта-функцию с центром в нуле. Как следствие, часть базисных функций (признаков) исключается (признается нерелевантными) из решающего правила, т.к. их веса обнуляются.

Согласно последовательному байесовскому подходу, прогноз метки класса t_{new} для тестового объекта $\boldsymbol{x}_{new}$ в RVM соответствует поиску распределения

$$p(t_{new}|\boldsymbol{x}_{new},T,X,\boldsymbol{\alpha}) = \int p(t_{new}|\boldsymbol{x}_{new},\boldsymbol{w})p(\boldsymbol{w}|T,X,\boldsymbol{\alpha})d\boldsymbol{w}.$$

По аналогии с обоснованностью (19) данный интеграл не вычисляется аналитически. Простейшей схемой аппроксимации здесь является замена апостериорного распределения $p(\boldsymbol{w}|T,X,\boldsymbol{\alpha})$ на дельта-функцию $\delta(\boldsymbol{w}-\boldsymbol{w}_{MP}(\boldsymbol{\alpha}))$, где $\boldsymbol{w}_{MP}(\boldsymbol{\alpha}) = \arg\max_{\boldsymbol{w}} Q(\boldsymbol{w},\boldsymbol{\alpha})$. Тогда

$$p(t_{new}|\boldsymbol{x}_{new},T,X,\boldsymbol{\alpha}) \simeq p(t_{new}|\boldsymbol{x}_{new},\boldsymbol{w}_{MP}(\boldsymbol{\alpha})).$$

Более точной схемой аппроксимации для прогнозного распределения является использование приближения Лапласа. Пример применения метода релевантных векторов показан на рис. 4e.

Проведем модельный эксперимент. Обучающая выборка состоит из 100 объектов (по 50 объектов из каждого класса) и пятидесяти признаков. При этом только два признака (пятый и семнадцатый) являются информативными с точки зрения разделения двух классов (см. рис. 5a). Остальные признаки являются шумовыми. Тестовая выборка сгенерирована аналогичным образом и состоит из 1000 объектов (по 500 объектов из каждого класса), см. рис. 5b. При обучении логистической регрессии все веса в линейном решающем правиле оказываются отличными от нуля (см. рис. 5c). Ошибка на тестовой выборке при этом составляет 16%. При обучении метода релевантных векторов только три веса оказываются отличными от нуля (см. рис. 5d), в том числе веса для пятого и семнадцатого признака. В результате ошибка на тестовой выборке сокращается до 8%, что соответствует байесовскому уровню ошибки для данной задачи.

Рассмотрим теперь метод RVM для случая K классов. Для этого модифицируем вероятностную модель мультиномиальной регрессии следующим образом:

$$p(T,W,\boldsymbol{\alpha}|X) = p(T|X,W)p(W|\boldsymbol{\alpha}) = p(W|\boldsymbol{\alpha})\prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n,W),$$

$$p(t_n|\boldsymbol{x}_n,W) = \frac{\exp(f_{t_n}(\boldsymbol{x}_n,\boldsymbol{w}^{t_n}))}{\sum_{j=1}^{K}\exp(f(\boldsymbol{x}_n,\boldsymbol{w}^j))},$$

$$p(W|\boldsymbol{\alpha}) = \prod_{j,m=1}^{K,M} \mathcal{N}(w_m^j|0,\alpha_m^{-1}).$$

Заметим, что в этой модели скалярный параметр регуляризации α_j вводится для всех весов из j-го столбца матрицы W. В результате устремление α_j к плюс бесконечности соответствует обнулению

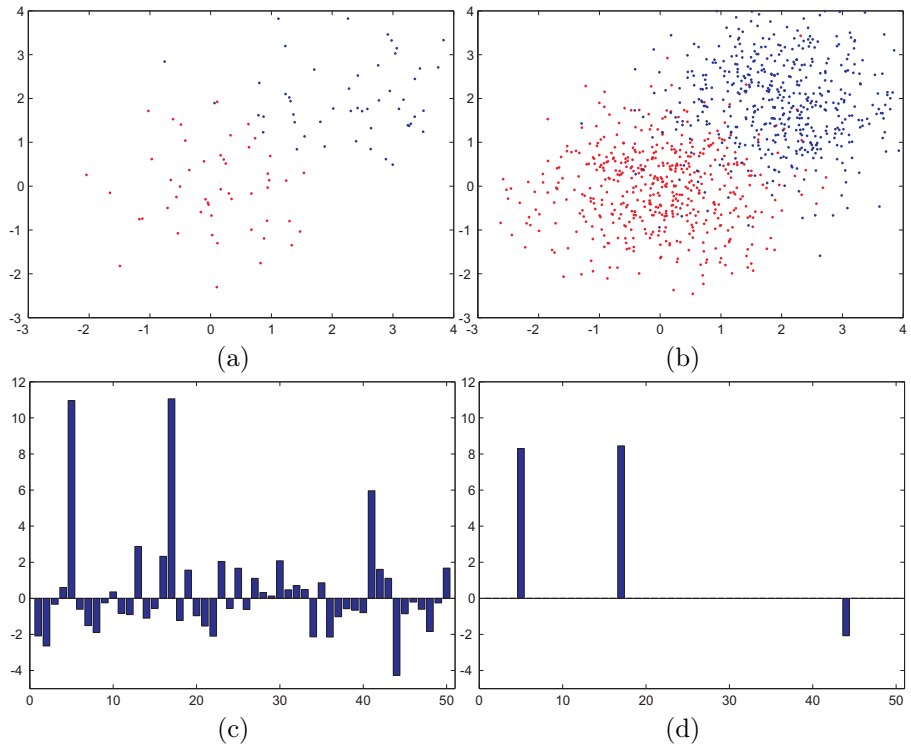

Рис. 5: Модельный эксперимент. (a) – два информативных признака в обучающей выборке, (b) – два информативных признака в тестовой выборке, (c) – веса обученной логистической регрессии, (d) – веса обученного метода релевантных векторов.

весов j-ой базисной функции сразу для всех линейных функций $f_1, \ldots, f_K$, т.е. исключению данного признака из рассмотрения.

Схема обучения и принятия решения в модели RVM для многих классов полностью повторяет схему для двух классов. На этапе обучения вектор $\boldsymbol{\alpha}$ настраивается путем максимизации обоснованности $p(T|X, \boldsymbol{\alpha})$, в которой используется приближение Лапласа. На этапе принятия решения для нового объекта в простейшем случае можно ограничиться вычислением весов $W_{MP}(\boldsymbol{\alpha})$ и применением решающего правила (17) для этих весов. Пример применения многоклассового метода релевантных векторов показан на рис. 4j.