Семинары по байесовским методам

Евгений Соколов
sokolov.evg@gmail.com

15 декабря 2016 г.

1 Байесовские методы машинного обучения

Пусть $X = \{x_1, \ldots, x_\ell\}$ — выборка, $\mathbb{X}$ — множество всех возможных объектов, Y — множество ответов. В байесовском подходе предполагается, что обучающие объекты и ответы на них $(x_1, y_1), \ldots, (x_\ell, y_\ell)$ независимо выбираются из некоторого распределения $p(x, y)$, заданного на множестве $\mathbb{X} \times Y$. Данное распределение можно переписать как

$$p(x, y) = p(y)p(x \mid y),$$

где $p(y)$ определяет вероятности появления каждого из возможных ответов и называется *априорным распределением*, а $p(x \mid y)$ задает распределение объектов при фиксированном ответе y и называется *функцией правдоподобия*.

Если известны априорное распределение и функция правдоподобия, то по формуле Байеса можно записать *апостериорное распределение* на множестве ответов:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{\int_s p(x \mid s)p(s)ds} = \frac{p(x \mid y)p(y)}{p(x)},$$

где знаменатель не зависит от y и является нормировочной константой.

§1.1 Оптимальные байесовские правила

Пусть на множестве всех пар ответов $Y \times Y$ задана функция потерь $L(y, s)$. Наиболее распространенным примером для задач классификации является ошибка классификации $L(y, s) = [y \neq s]$, для задач регрессии — квадратичная функция потерь $L(y, x) = (y - s)^2$. *Функционалом среднего риска* называется матожидание функции потерь по всем парам (x, y) при использовании алгоритма $a(x)$:

$$R(a) = \mathbb{E}L(y, a(x)) = \int_Y \int_{\mathbb{X}} L(y, a(x))p(x, y)dxdy.$$

Если распределение $p(x, y)$ известно, то можно найти алгоритм $a_*(x)$, оптимальный с точки зрения функционала среднего риска.

1.1.1 Классификация

Начнем с задачи классификации с множеством ответом $Y = \{1, \dots, K\}$ и функции потерь $L(y, s) = [y \neq s]$. Покажем, что минимум функционала среднего риска достигается на алгоритме

$$a_*(x) = \arg\max_{y \in Y} p(y \mid x).$$

Для произвольного классификатора $a(x)$ выполнена следующая цепочка неравенств [1]:

$$\begin{aligned}
R(a) &= \int_Y \int_{\mathbb{X}} L(y, a(x)) p(x, y) dx dy = \\
&= \sum_{y=1}^{K} \int_{\mathbb{X}} [y \neq a(x)] p(x, y) dx = \\
&= \int_{\mathbb{X}} \sum_{y \neq a(x)} p(x, y) dx = \left\{ \int_{\mathbb{X}} \sum_{y \neq a(x)} p(x, y) dx + \int_{\mathbb{X}} p(x, a(x)) dx = 1 \right\} = \\
&= 1 - \int_{\mathbb{X}} p(x, a(x)) dx \geqslant \\
&\geqslant 1 - \int_{\mathbb{X}} \max_{s \in Y} p(x, s) dx = \\
&= 1 - \int_{\mathbb{X}} p(x, a_*(x)) dx = \\
&= R(a_*)
\end{aligned}$$

Таким образом, средний риск любого классификатора $a(x)$ не превосходит средний риск нашего классификатора $a_*(x)$.

Мы получили, что оптимальный байесовский классификатор выбирает тот класс, который имеет наибольшую апостериорную вероятность. Такой классификатор называется *MAP-классификатором* (maximum a posteriori).

1.1.2 Регрессия

Перейдем к задаче регрессии и функции потерь $L(y, x) = (y - s)^2$. Нам пригодится понятие условного матожидания:

$$\mathbb{E}(y \mid x) = \int_Y y p(y \mid x) dy.$$

Преобразуем функцию потерь [2]:

$$\begin{aligned}
L(y, a(x)) = (y - a(x))^2 &= (y - \mathbb{E}(y \mid x) + \mathbb{E}(y \mid x) - a(x))^2 = \\
&= (y - \mathbb{E}(y \mid x))^2 + 2\big(y - \mathbb{E}(y \mid x)\big)\big(\mathbb{E}(y \mid x) - a(x)\big) + (\mathbb{E}(y \mid x) - a(x))^2.
\end{aligned}$$

Подставляя ее в функционал среднего риска, получаем:

$$\begin{aligned}R(a) &= \int_Y \int_{\mathbb{X}} L(y, a(x))p(x, y)dxdy = \\ &= \int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t \mid x))^2 p(x, y)dxdy + \int_Y \int_{\mathbb{X}} (\mathbb{E}(t \mid x) - a(x))^2 p(x, y)dxdy + \\ &+ 2\int_Y \int_{\mathbb{X}} \big(y - \mathbb{E}(t \mid x)\big)\big(\mathbb{E}(t \mid x) - a(x)\big)p(x, y)dxdy.\end{aligned}$$

Разберемся сначала с последним слагаемым. Заметим, что величина $\big(\mathbb{E}(t \mid x) - a(x)\big)$ не зависит от y, и поэтому ее можно вынести за интеграл по y:

$$\begin{aligned}&\int_Y \int_{\mathbb{X}} \big(y - \mathbb{E}(t \mid x)\big)\big(\mathbb{E}(t \mid x) - a(x)\big)p(x, y)dxdy = \\ &= \int_{\mathbb{X}} \big(\mathbb{E}(t \mid x) - a(x)\big) \int_Y \big\{\big(y - \mathbb{E}(t \mid x)\big)p(x, y)\big\}\, dydx = \\ &= \int_{\mathbb{X}} \big(\mathbb{E}(t \mid x) - a(x)\big) \left\{\int_Y yp(x, y)dy - \int_Y \mathbb{E}(t \mid x)p(x, y)dy\right\} dx = \\ &= \int_{\mathbb{X}} \big(\mathbb{E}(t \mid x) - a(x)\big) \left\{p(x)\int_Y yp(y \mid x)dy - \mathbb{E}(t \mid x)\int_Y p(x, y)dy\right\} dx = \\ &= \int_{\mathbb{X}} \big(\mathbb{E}(t \mid x) - a(x)\big) \underbrace{\big\{p(x)\mathbb{E}(t \mid x) - p(x)\mathbb{E}(t \mid x)\big\}}_{=0} dx = \\ &= 0\end{aligned}$$

Получаем, что функционал среднего риска имеет вид

$$R(a) = \int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t \mid x))^2 p(x, y)dxdy + \int_Y \int_{\mathbb{X}} (\mathbb{E}(t \mid x) - a(x))^2 p(x, y)dxdy.$$

От алгоритма $a(x)$ зависит только второе слагаемое, и оно достигает своего минимума, если $a(x) = \mathbb{E}(t \mid x)$. Таким образом, оптимальная байесовская функция регрессии для квадратичной функции потерь имеет вид

$$a_*(x) = \mathbb{E}(y \mid x) = \int_Y yp(y \mid x)dy.$$

Иными словами, мы должны провести «взвешенное голосование» по всем возможным ответам, причем вес ответа равен его апостериорной вероятности.

§1.2 Байесовский вывод

Основной проблемой оптимальных байесовских алгоритмов, о которых шла речь в предыдущем разделе, является невозможность их построения на практике, поскольку нам никогда неизвестно распределение $p(x, y)$. Данное распределение можно попробовать восстановить по обучающей выборке, при этом существует два подхода — параметрический и непараметрический. Сейчас мы сосредоточимся на параметрическом подходе.

Допустим, распределение на парах «объект-ответ» зависит от некоторого параметра θ: $p(x, y \mid \theta)$. Тогда получаем следующую формулу для апостериорной вероятности:

$$p(y \mid x, \theta) \propto p(x \mid y, \theta)p(y),$$

где выражение «$a \propto b$» означает «a пропорционально b». Для оценивания параметров применяется *метод максимального правдоподобия*:

$$\theta_* = \arg\max_\theta L(\theta) = \arg\max_\theta \prod_{i=1}^{\ell} p(x_i \mid y_i, \theta),$$

где $L(\theta)$ — функция правдоподобия. Примером такого подхода может служить *нормальный дискриминантный анализ*, где предполагается, что функции правдоподобия являются нормальными распределениями с неизвестными параметрами $\theta = (\mu, \Sigma)$.

Иногда удобнее сразу задавать апостериорное распределение — например, в случае с линейной регрессией. Будем считать, что задан некоторый вектор весов w, и метка объекта $y(x)$ генерируется следующим образом: вычисляется линейная функция $\langle w, x\rangle$, и к результату прибавляется нормальный шум:

$$y(x) = \langle w, x\rangle + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

В этом случае апостериорное распределение примет вид

$$p(y \mid x, w) = \mathcal{N}(\langle w, x\rangle, \sigma^2). \tag{1.1}$$

Задача 1.1. *Покажите, что метод максимального правдоподобия для модели* (1.1) *эквивалентен методу наименьших квадратов.*

Решение. Запишем правдоподобие для выборки $x_1, \ldots, x_\ell$:

$$L(w) = \prod_{i=1}^{\ell} p(y_i \mid x_i, w) = \prod_{i=1}^{\ell} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle w, x_i\rangle)^2}{2\sigma^2}\right).$$

Перейдем к логарифму правдоподобия:

$$\log L(w) = -\ell \log \sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{\ell} (y_i - \langle w, x_i\rangle)^2 \to \max_w.$$

Убирая все члены, не зависящие от вектора весов w, получаем задачу наименьших квадратов

$$\sum_{i=1}^{\ell} (y_i - \langle w, x_i\rangle)^2 \to \min_w.$$

■

Байесовский вывод параметров. В некоторых случаях применение метода максимального правдоподобия для поиска параметров приводит к плохим результатам. Например, если имеет место мультиколлинеарность, то функция правдоподобия имеет много минимумов, и решение может оказаться переобученным. Одним из подходов к устранению этой проблемы является введение априорного распределения *на параметрах*.

Пусть $p(\theta)$ — априорное распределение на векторе параметров θ. В качестве функции правдоподобия для данного вектора возьмем апостериорное распределение на ответах $p(y \mid x, \theta)$. Тогда по формуле Байеса

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)p(\theta)}{p(y \mid x)}.$$

Вернемся к примеру с линейной регрессией. Введем априорное распределение на векторе весов:

$$p(w_j) = \mathcal{N}(0, \alpha^2), \quad j = 1, \ldots, d.$$

Иными словами, мы предполагаем, что веса концентрируются вокруг нуля.

Задача 1.2. *Покажите, что максимизация апостериорной вероятности $p(w \mid y, x)$ для модели линейной регрессии с нормальным априорным распределением эквивалентна решению задачи гребневой регрессии.*

Решение. Запишем апостериорную вероятность вектора весов w для выборки $x_1, \ldots, x_\ell$:

$$p(w \mid y, x) = \prod_{i=1}^{\ell} p(y_i \mid x_i, w)p(w) =$$
$$= \prod_{i=1}^{\ell} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle w, x_i\rangle)^2}{2\sigma^2}\right) \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{w_j^2}{2\alpha^2}\right).$$

Перейдем к логарифму и избавимся от константных членов:

$$\log p(w \mid y, x) = -\frac{1}{2\sigma^2} \sum_{i=1}^{\ell} (y_i - \langle w, x_i\rangle)^2 - \frac{\ell}{2\alpha^2} \underbrace{\sum_{j=1}^{d} w_j^2}_{=\|w\|^2}.$$

В итоге получаем задачу гребневой регрессии

$$\sum_{i=1}^{\ell} (y_i - \langle w, x_i\rangle)^2 + \lambda\|w\|^2 \to \min_w,$$

где $\lambda = \frac{\ell}{2\alpha^2}$.

■

После того, как оптимальный вектор весов w_* найден, мы можем найти распределение на ответах для нового объекта x:

$$p(y \mid x, X, w_*) = \mathcal{N}(\langle x, w_*\rangle, \sigma^2).$$

Выше мы выяснили, что оптимальным ответом будет матожидание $\mathbb{E}(y \mid x) = \int yp(y \mid x, X, w_*)dy$.

С точки зрения байесовского подхода [3] правильнее не искать моду [1] w_* апостериорного распределения на параметрах и брать соответствующую ей модель $p(y \mid x, X, w_*)$, а устроить «взвешенное голосование» всех возможных моделей:

$$p(y \mid x, X) = \int p(y \mid x, w)p(w \mid Y, X)dw,$$

где $X = \{x_1, \ldots, x_\ell\}$, $Y = \{y_1, \ldots, y_\ell\}$.

2 Нормальный дискриминантный анализ

Нормальный дискриминантный анализ — это частный случай байесовской классификации, когда предполагается, что функции правдоподобия классов $p(x \mid y)$ являются нормальными.

§2.1 Векторное дифференцирование

Выведем формулы векторного дифференцирования, которые пригодятся нам при работе с плотностями нормальных распределений.

Задача 2.1. *Покажите, что*

$$\nabla_X a^T X b = ab^T,$$

где $a \in \mathbb{R}^m$, $b \in \mathbb{R}^n$, $X \in \mathbb{R}^{m \times n}$.

Решение. Вспомним, что производная по матрице — это матрица частных производных по компонентам этой матрицы. Найдем их:

$$\frac{\partial}{\partial x_{ij}} a^T X b = \frac{\partial}{\partial x_{ij}} \sum_{k=1}^{m} \sum_{s=1}^{n} x_{ks} a_k b_s = a_i b_j.$$

Таким образом,

$$\nabla_X a^T X b = ab^T = (a_i b_j)_{i,j} = ab^T.$$

■

Задача 2.2. *Покажите, что*

$$\nabla_X \log \det X = X^{-T},$$

где $X \in \mathbb{R}^{n \times n}$ *— положительно определенная матрица* [2].

[1] Мода — точка максимума плотности.

[2] Если матрица не положительно определена, то ее определитель может быть отрицательным или равным нулю, и логарифм от него будет неопределен.

Решение. Запишем производную по x_{ij}:

$$\frac{\partial}{\partial x_{ij}} \log\det X = \frac{1}{\det X}\frac{\partial \det X}{\partial x_{ij}}.$$

Вспомним *теорему Лапласа* из линейной алгебры и несколько связанных с ней определений. *Минором* M_{ij} матрицы X называется определитель матрицы, полученной из X вычеркиванием i-й строки и j-го столбца [3]. *Алгебраическим дополнением* C_{ij} матрицы X называется величина $(-1)^{i+j}M_{ij}$. Теорема Лапласа гласит, что определитель матрицы X можно выразить через ее алгебраические дополнения:

$$\det X = \sum_{k=1}^{n} x_{kj}C_{kj}. \tag{2.1}$$

Вернемся к вычислению производной $\partial \det X/\partial x_{ij}$. Заметим, что в разложении (2.1) все алгебраические дополнения вычисляются по матрицам, в которых отсутствует элемент x_{ij}, и поэтому они могут быть вынесены за знак производной. Получаем, что

$$\frac{\partial \det X}{\partial x_{ij}} = \frac{\partial}{\partial x_{ij}}\sum_{k=1}^{n} x_{kj}C_{kj} = \sum_{k=1}^{n} C_{kj}\frac{\partial}{\partial x_{ij}}x_{kj} = C_{ij}.$$

Отсюда следует, что

$$\nabla_X \log\det X = \frac{1}{\det X}(C_{ij})_{i,j=1}^{n} = \frac{1}{\det X}(X^*)^T.$$

Матрица $X^* = (C_{ji})$, составленная из алгебраических дополнений к матрице X, называется *союзной* или *присоединенной*. Из линейной алгебры известно, что союзная матрица пропорциональна обратной:

$$X^{-1} = \frac{1}{\det X}X^*.$$

Учитывая это, получаем:

$$\nabla_X \log\det X = \frac{1}{\det X}(X^*)^T = \frac{1}{\det X}(X^{-1}\det X)^T = \frac{\det X}{\det X}X^{-T} = X^{-T}.$$

■

Задача 2.3. *Покажите, что*

$$\nabla_X \log\det X^{-1} = -X^{-T},$$

где $X \in \mathbb{R}^{n\times n}$ *— положительно определенная матрица.*

Решение.

$$\nabla_X \log\det X^{-1} = \nabla_X \log(\det X)^{-1} = -\nabla_X \log\det X = -X^{-T}.$$

■

[3] Строго говоря, минор - это определитель произвольной подматрицы, но здесь нам понадобятся миноры именно такого вида

§2.2 Нормальное распределение

Одномерное нормальное распределение. Случайная величина x имеет нормальное распределение, если ее плотность имеет вид

$$p(x \mid \mu, \sigma^2) = \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Вычисляя соответствующие интегралы, можно показать, что параметры μ и σ^2 соответствуют матожиданию и дисперсии:

$$\mathbb{E}x = \mu;$$
$$\mathbb{D}x = \sigma^2.$$

Центральная предельная теорема гласит, что среднее арифметическое независимых одинаково распределенных случайных величин стремится к нормальному распределению:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} \mathcal{N}(x \mid 0, \sigma^2),$$

где μ и σ^2 — матожидание и дисперсия данных случайных величин.

Известно, что нормальное распределение имеет легкие хвосты — вероятность того, что нормальная случайная величина отклонится от своего среднего больше, чем на 3σ, не превышает 0.3%. Этот факт называют «правилом трех сигм».

Многомерное нормальное распределение. Случайный вектор $x = (x_1, \ldots, x_d)$ имеет многомерное нормальное распределение, если его плотность имеет вид

$$p(x \mid \mu, \Sigma) = \mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}(\det\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

Матрица Σ должна быть симметричной и положительно определенной.

Вычисляя соответствующие интегралы, можно показать, что параметры μ и Σ соответствуют матожиданию и ковариационной матрице:

$$\mathbb{E}x = \mu;$$
$$\mathbb{E}(x-\mu)(x-\mu)^T = \Sigma;$$
$$\mathbb{D}x_i = \Sigma_{ii};$$
$$\mathrm{Cov}(x_i, x_j) = \mathbb{E}(x_i - \mu_i)(x_j - \mu_j) = \Sigma_{ij}.$$

Можно показать, что все моменты многомерной случайной величины выражаются через среднее μ и ковариационную матрицу Σ.

Существует обобщение центральной предельной теоремы на многомерный случай, которое гласит, что среднее арифметическое независимых одинаково распределенных случайных векторов стремится к многомерному нормальному распределению:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} \mathcal{N}(x \mid 0, \Sigma),$$

где μ и Σ — матожидание и ковариационная матрица случайных величин.

Линии уровня плотности нормального распределения соответствуют линиям уровня квадратичной формы $(x-\mu)^T\Sigma^{-1}(x-\mu)$ и представляют собой эллипсы. Ранее мы подробно выводили вид линий уровня таких квадратичных форм, когда сталкивались с расстоянием Махалонобиса (см. семинары по метрическим методам).

§2.3 Нормальный дискриминантный анализ

Оптимальный байесовский классификатор при бинарной функции потерь имеет вид

$$a(x) = \arg\max_{y\in Y} p(y)p(x\,|\,y).$$

В нормальном дискриминантном анализе предполагается, что распределения объектов внутри классов $p(x\,|\,y)$ — нормальные:

$$p(x\,|\,y) = \mathcal{N}(x\,|\,\mu_y, \Sigma_y).$$

Параметрами алгоритма являются средние μ_y и ковариационные матрицы классов Σ_y, которые оцениваются по выборке методом максимального правдоподобия.

Задача 2.4. *Выведите оценку максимального правдоподобия на вектор матожиданий μ_y, если к классу y относятся объекты выборки $X_y = \{x_1, \ldots, x_m\}$.*

Решение. Для краткости будем обозначать вектор матожиданий и ковариационную матрицу для класса y через μ и Σ. Нам нужно решить задачу

$$p(X_y\,|\,\mu,\Sigma) = \prod_{i=1}^{m}\mathcal{N}(x_i\,|\,\mu,\Sigma) \to \max_{\mu}.$$

Перейдем к логарифму:

$$\log p(X_y\,|\,\mu,\Sigma) = -\frac{m}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) + \text{const}.$$

Найдем производную по μ и приравняем ее к нулю:

$$\begin{aligned}
\nabla_\mu \log p(X_y\,|\,\mu,\Sigma) &= -\frac{1}{2}\nabla_\mu\left(\sum_{i=1}^{m}x_i^T\Sigma^{-1}x_i - 2\sum_{i=1}^{m}x_i^T\Sigma^{-1}\mu + \sum_{i=1}^{m}\mu^T\Sigma^{-1}\mu\right) = \\
&= -\frac{1}{2}\left(-2\sum_{i=1}^{m}\underbrace{\Sigma^{-T}}_{=\Sigma^{-1}}x_i + \sum_{i=1}^{m}2\Sigma^{-1}\mu\right) = \\
&= \Sigma^{-1}\left(m\mu - \sum_{i=1}^{m}x_i\right) = \\
&= 0.
\end{aligned}$$

Домножая слева на матрицу Σ, получаем

$$\mu = \frac{1}{m}\sum_{i=1}^{m}x_i.$$

■

Задача 2.5. *Выведите оценку максимального правдоподобия на ковариационную матрицу Σ, если к классу y относятся объекты выборки $X_y = \{x_1, \ldots, x_m\}$.*

Решение. Как и в предыдущей задачи, будем обозначать вектор матожиданий и ковариационную матрицу для класса y через μ и Σ.

Для удобства перейдем в правдоподобии к матрице точности $\Lambda = \Sigma^{-1}$:

$$\log p(X_y \,|\, \mu, \Lambda) = -\frac{m}{2} \log \det \Lambda^{-1} - \frac{1}{2} \sum_{i=1}^{m} (x_i - \mu)^T \Lambda (x_i - \mu) + \text{const}.$$

Найдем производную по Λ и приравняем ее к нулю:

$$\begin{aligned}
\nabla_\Lambda \log p(X_y \,|\, \mu, \Lambda) &= -\frac{m}{2} \nabla_\Sigma \log \det \Lambda^{-1} - \frac{1}{2} \sum_{i=1}^{m} \nabla_\Lambda (x_i - \mu)^T \Lambda (x_i - \mu) = \\
&= \frac{m}{2} \underbrace{\Lambda^{-T}}_{=\Lambda^{-1}} - \frac{1}{2} \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T = \\
&= 0
\end{aligned}$$

Отсюда

$$\Lambda = \frac{1}{m} \left(\sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T \right)^{-1}.$$

Переходя обратно к ковариационной матрице Σ, получаем

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^T.$$

■

2.3.1 Линейный дискриминант Фишера

Если предположить, что ковариационные матрицы классов равны, и оценивать их по всей выборке, то мы получим алгоритм, называемый *линейным дискриминантом Фишера*. Можно показать, что он является линейным:

$$a(x) = \arg\max_{y \in Y} (\langle w_y, x \rangle + w_{0y}),$$

причем $w_y = \Sigma^{-1} \mu_y$. В случае двух классов ($Y = \{-1, +1\}$) классификатор принимает вид

$$a(x) = \text{sign}\,(\langle w, x \rangle + b) \quad w = \Sigma^{-1}(\mu_2 - \mu_1). \tag{2.2}$$

Разберем другую интерпретацию линейного дискриминанта Фишера. Будем классифицировать объекты следующим образом: выберем прямую с направляющим вектором w и спроецируем объект на нее; если значение проекции окажется больше порога $-b$, то отнесем объект к классу $+1$, иначе к классу -1. Таким образом, классификатор будет иметь вид $a(x) = \text{sign}\,(\langle w, x \rangle + b)$. Обучение классификатора сводится

к поиску проекционной прямой. Будем выбирать ее так, чтобы после проецирования разброс точек из одного класса был как можно меньше, а расстояние между центрами классов было как можно больше. Формализуем эти требования. Обозначим через m_k центр k-го класса, $k \in Y$:

$$m_k = \frac{1}{N_k} \sum_{i:y_i=k} x_i.$$

Пусть s_k^2 — внутриклассовая дисперсия класса k:

$$s_k^2 = \sum_{i:y_i=k} (w^T x_i - w^T m_k)^2.$$

В качестве меры «сгруппированности» точек внутри своих классов возьмем сумму внутриклассовых дисперсий $s_{-1}^2 + s_{+1}^2$. В качестве меры расстояния между центрами проекций классов («межклассовой дисперсии») возьмем квадрат расстояния между этими центрами: $(w^T m_{-1} - w^T m_{+1})^2$. Чтобы совместить минимизацию первой величины и максимизацию второй, возьмем в качестве функционала их отношение. Получим следующую оптимизационную задачу:

$$J(w) = \frac{(w^T m_{-1} - w^T m_{+1})^2}{s_{-1}^2 + s_{+1}^2} \to \max_w.$$

Распишем данный функционал:

$$\begin{aligned} J(w) &= \frac{(w^T m_{-1} - w^T m_{+1})^2}{s_{-1}^2 + s_{+1}^2} = \\ &= \frac{\left(w^T (m_{-1} - m_{+1})\right)^2}{\sum_{i:y_i=-1} \left(w^T (x_i - m_{-1})\right)^2 + \sum_{i:y_i=+1} \left(w^T (x_i - m_{+1})\right)^2} = \\ &= \frac{w^T (m_{-1} - m_{+1})(m_{-1} - m_{+1})^T w}{\sum_{i:y_i=-1} w^T (x_i - m_{-1})(x_i - m_{-1})^T w + \sum_{i:y_i=+1} w^T (x_i - m_{+1})(x_i - m_{+1})^T w} = \\ &= \frac{w^T (m_{-1} - m_{+1})(m_{-1} - m_{+1})^T w}{w^T \left(\sum_{i:y_i=-1} (x_i - m_{-1})(x_i - m_{-1})^T + \sum_{i:y_i=+1} (x_i - m_{+1})(x_i - m_{+1})^T \right) w}. \end{aligned}$$

Введем обозначения для ковариационных матриц:

$$\begin{aligned} S_b &= (m_{-1} - m_{+1})(m_{-1} - m_{+1})^T; \\ S_w &= \sum_{i:y_i=-1} (x_i - m_{-1})(x_i - m_{-1})^T + \sum_{i:y_i=+1} (x_i - m_{+1})(x_i - m_{+1})^T. \end{aligned}$$

Тогда функционал примет вид

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \to \max_w.$$

Нам понадобится следующее правило векторного дифференцирования.

Задача 2.6. *Покажите, что если* $f : \mathbb{R}^d \to \mathbb{R}$ *и* $g : \mathbb{R}^d \to \mathbb{R}$ *— вещественные функции, то*

$$\nabla_x \frac{f(x)}{g(x)} = \frac{g(x)\nabla_x f(x) - f(x)\nabla_x g(x)}{g^2(x)}.$$

Воспользуемся полученным правилом, чтобы вычислить градиент функционала $J(w)$ и приравнять его нулю:

$$\begin{aligned}\nabla_w J(w) &= \frac{(S_b + S_b^T)w(w^T S_w w) - (S_w + S_w^T)w(w^T S_b w)}{(w^T S_w w)^2} = \\ &= 2\frac{S_b w(w^T S_w w) - S_w w(w^T S_b w)}{(w^T S_w w)^2} = \\ &= 0.\end{aligned}$$

Приходим к уравнению

$$S_b w(w^T S_w w) = S_w w(w^T S_b w). \tag{2.3}$$

Пусть минимум функционала $J(w)$ достигается на векторе w_*. Тогда этот вектор удовлетворяет уравнению (2.3). Поскольку классификатор (2.2) зависит только от направления вектора w и не зависит от его длины, мы можем проигнорировать скалярные множители. Получаем:

$$\begin{aligned}S_w w_* &= \\ &= \underbrace{\frac{w_*^T S_w w_*}{w_*^T S_b w_*}}_{\in\mathbb{R}} S_b w_* \propto \\ &\propto S_b w_* = \\ &= (m_{-1} - m_{+1})\underbrace{(m_{-1} - m_{+1})^T w_*}_{\in\mathbb{R}} \propto \\ &\propto (m_{-1} - m_{+1}).\end{aligned}$$

Значит,

$$w_* = S_w^{-1}(m_{-1} - m_{+1}).$$

Мы пришли к такому же вектору весов w, который может быть получен при нормальном дискриминантном анализе в предположении о равенстве ковариационных матриц классов.

Список литературы

[1] *Ветров, Д.П., Кропотов, Д.А.* Байесовские методы машинного обучения. Учебное пособие. // Москва, 2007.

[2] *Bishop, C.M.* Pattern Recognition and Machine Learning. // Springer, 2006.

[3] *Murphy, K.P.* Machine Learning: A Probabilistic Perspective. // MIT Press, 2012.