

Методы машинного обучения

Вероятностные модели порождения данных

Воронцов Константин Вячеславович

www.MachineLearning.ru/wiki?title=User:Vokov

вопросы к лектору: k.vorontsov@iai.msu.ru

материалы курса:

github.com/MSU-ML-COURSE/ML-COURSE-24-25

орг.вопросы по курсу: ml.cmc@mail.ru

1 Принцип максимума правдоподобия

- Задача оценивания плотности распределения
- Задача обучения регрессии
- Задача обучения классификации

2 Логистическая регрессия

- Вероятностный линейный классификатор
- Предсказание вероятности класса
- Вероятностная калибровка Платта

3 Основы байесовской теории классификации

- Оптимальный байесовский классификатор
- Наивный байесовский классификатор
- Нормальный дискриминантный анализ

Задача восстановления плотности — обучение без учителя

Дано: простая (i.i.d.) выборка $X^\ell = \{x_1, \dots, x_\ell\} \sim p(x)$

Найти параметрическую модель плотности распределения:

$$p(x) = \varphi(x; w),$$

где w — вектор параметров, φ — фиксированная функция

Критерий — максимум (логарифма) правдоподобия выборки,
MLE-оценивание параметра w (Maximum Likelihood Estimate):

$$L(w; X^\ell) = \ln \prod_{i=1}^{\ell} \varphi(x_i; w) = \sum_{i=1}^{\ell} \ln \varphi(x_i; w) \rightarrow \max_w$$

Аналитическое решение: необходимое условие экстремума

$$\frac{\partial}{\partial w} L(w; X^\ell) = \sum_{i=1}^{\ell} \frac{\partial}{\partial w} \ln \varphi(x_i; w) = 0,$$

при условии достаточной гладкости функции $\varphi(x; w)$ по w

Частный случай: многомерная гауссовская плотность

Пусть объекты x описываются n признаками $f_j(x) \in \mathbb{R}$
 и выборка порождена n -мерной гауссовской плотностью:

$$x_i \sim p(x) = \mathcal{N}(x; \mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)}{\sqrt{(2\pi)^n \det \Sigma}}, \quad x \in \mathbb{R}^n$$

$\mu \in \mathbb{R}^n$ — вектор математического ожидания, $\mu = \mathbb{E}x$
 $\Sigma \in \mathbb{R}^{n \times n}$ — ковариационная матрица, $\Sigma = \mathbb{E}(x - \mu)(x - \mu)^\top$
 (симметричная, невырожденная, положительно определённая)

Выборочные оценки максимального правдоподобия:

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln L(\mu, \Sigma; X^\ell) = 0 & \Rightarrow \hat{\mu} = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i \\ \frac{\partial}{\partial \Sigma} \ln L(\mu, \Sigma; X^\ell) = 0 & \Rightarrow \hat{\Sigma} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \hat{\mu})(x_i - \hat{\mu})^\top \end{aligned}$$

Некоторые приёмы матричного дифференцирования

Производная скалярной функции $f(A)$ по матрице $A = (a_{ij})$:

$$\frac{\partial}{\partial A} f(A) = \left(\frac{\partial}{\partial a_{ij}} f(A) \right)$$

$\text{diag } A$ — диагональ матрицы A , остальные элементы нули

A — квадратная $n \times n$ -матрица

u — вектор размерности n

если A произвольного вида:

$$\frac{\partial}{\partial u} u^T A u = A^T u + A u$$

$$\frac{\partial}{\partial A} \ln |A| = A^{-1T}$$

$$\frac{\partial}{\partial A} u^T A u = u u^T$$

если A симметричная, $A^T = A$:

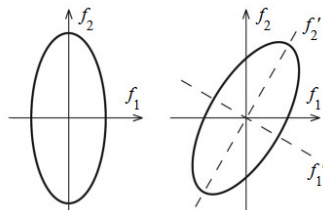
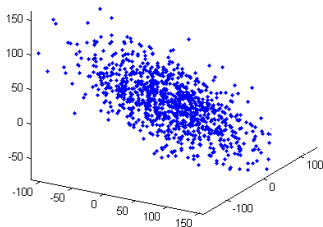
$$\frac{\partial}{\partial u} u^T A u = 2A u$$

$$\frac{\partial}{\partial A} \ln |A| = 2A^{-1} - \text{diag } A^{-1}$$

$$\frac{\partial}{\partial A} u^T A u = 2u u^T - \text{diag } u u^T$$

Геометрический смысл многомерной нормальной плотности

Эллипсоид рассеяния — облако точек эллиптической формы:



При $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ оси эллипсоида параллельны осям.

В общем случае: $\Sigma = VSV^T$ — спектральное разложение,

$V = (v_1, \dots, v_n)$ — ортогональные собств. векторы, $V^T V = I_n$

$S = \text{diag}(\lambda_1, \dots, \lambda_n)$ — собственные значения матрицы Σ

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = (x - \mu)^T V S^{-1} V^T (x - \mu) = (x' - \mu')^T S^{-1} (x' - \mu').$$

$x' = V^T x$ — ортогональное преобразование поворот/отражение

Проблема мультиколлинеарности

Проблема: при $\ell < n$ матрица $\hat{\Sigma}$ вырождена, но даже при $\ell \geq n$ она может оказаться плохо обусловленной.

Регуляризация ковариационной матрицы $\hat{\Sigma} + \tau I_n$ увеличивает собственные значения на τ , сохраняя собственные векторы (параметр τ можно подбирать по скользящему контролю)

Диагонализация ковариационной матрицы:

«наивное» предположение о независимости признаков приводит к оцениванию n одномерных плотностей признаков:

$$\hat{p}_j(\xi) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} \exp\left(-\frac{(\xi - \hat{\mu}_j)^2}{2\hat{\sigma}_j^2}\right), \quad j = 1, \dots, n$$

$\hat{\mu}_j = \frac{1}{\ell} \sum_{i=1}^{\ell} f_j(x_i)$ — выборочная оценка среднего признака f_j

$\hat{\sigma}_j^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} (f_j(x_i) - \hat{\mu}_j)^2$ — выборочная оценка дисперсии f_j

Задача регрессии и принцип максимума правдоподобия

Модель данных с некоррелированным гауссовским шумом:

$$y(x_i) = f(x_i, w) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad i = 1, \dots, \ell.$$

Метод максимума правдоподобия (ММП, MLE):

$$L(\varepsilon_1, \dots, \varepsilon_\ell; w) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2} \varepsilon_i^2\right) \rightarrow \max_w;$$

$$-\ln L(\varepsilon_1, \dots, \varepsilon_\ell; w) = \text{const}(w) + \frac{1}{2} \sum_{i=1}^{\ell} \frac{1}{\sigma_i^2} (f(x_i, w) - y_i)^2 \rightarrow \min_w;$$

Теорема

Постановки МНК и ММП, совпадают, причём веса объектов обратно пропорциональны дисперсии шума, $w_i = \sigma_i^{-2}$.

Задача классификации и принцип максимума правдоподобия

Пусть $X \times Y$ — в.п. с плотностью $p(x, y)$

Пусть X^ℓ — простая (i.i.d.) выборка: $(x_i, y_i)_{i=1}^\ell \sim p(x, y)$

Задача: по выборке X^ℓ оценить плотность $p(x, y)$

$p(x, y) = P(y|x, w)p(x)$ — параметризация плотности

$P(y|x, w)$ — модель условной вероятности класса с параметром w

$p(x)$ — неизвестное и непараметризуемое распределение на X

MLE-оценка для w (Maximum Likelihood Estimate):

$$\prod_{i=1}^{\ell} p(x_i, y_i) = \prod_{i=1}^{\ell} P(y_i|x_i, w) p(x_i) \rightarrow \max_w$$

Логарифм правдоподобия (log-likelihood, log-loss):

$$L(w) = \sum_{i=1}^{\ell} \log P(y_i|x_i, w) \rightarrow \max_w$$

Связь правдоподобия и аппроксимации эмпирического риска

Максимизация логарифма правдоподобия,

$P(y|x, w)$ — модель условной вероятности класса:

$$L(w) = \sum_{i=1}^{\ell} \log P(y_i|x_i, w) \rightarrow \max_w$$

Минимизация аппроксимированного эмпирического риска,

$g(x, w)$ — модель разделяющей поверхности, $Y = \{\pm 1\}$:

$$Q(w) = \sum_{i=1}^{\ell} \mathcal{L}(y_i g(x_i, w)) \rightarrow \min_w$$

Эти два принципа эквивалентны, если положить

$$-\log P(y_i|x_i, w) = \mathcal{L}(y_i g(x_i, w)).$$

$$\boxed{\text{модель } P(y|x, w)} \Leftrightarrow \boxed{\text{модель } g(x, w) \text{ и } \mathcal{L}(M)}.$$

Вероятностный смысл регуляризации

$P(y|x, w)$ — вероятностная модель данных

$p(w; \gamma)$ — априорное распределение параметров модели

γ — вектор гиперпараметров

Теперь не только появление выборки X^ℓ ,
 но и появление модели w также полагается стохастическим

Совместное правдоподобие данных и модели:

$$p(X^\ell, w) = p(X^\ell|w) \textcolor{red}{p(w; \gamma)}$$

Принцип максимума апостериорной вероятности

(Maximum a Posteriori Probability, MAP):

$$L(w) = \ln p(X^\ell, w) = \underbrace{\sum_{i=1}^{\ell} \log P(y_i|x_i, w)}_{\text{log правдоподобия}} + \underbrace{\textcolor{red}{\log p(w; \gamma)}}_{\substack{\text{регуляризатор,} \\ \text{не зависит от } X^\ell}} \rightarrow \max_w$$

Примеры: априорные распределения Гаусса и Лапласа

Пусть веса w_j независимы, $E w_j = 0$, $D w_j = C$.

Распределение Гаусса и квадратичный (L_2) регуляризатор:

$$p(w; C) = \frac{1}{(2\pi C)^{n/2}} \exp\left(-\frac{\|w\|^2}{2C}\right), \quad \|w\|^2 = \sum_{j=1}^n w_j^2,$$

$$-\ln p(w; C) = \frac{1}{2C} \|w\|^2 + \text{const}$$

Распределение Лапласа и абсолютный (L_1) регуляризатор:

$$p(w; C) = \frac{1}{(2C)^n} \exp\left(-\frac{\|w\|}{C}\right), \quad \|w\| = \sum_{j=1}^n |w_j|,$$

$$-\ln p(w; C) = \frac{1}{C} \|w\| + \text{const}$$

C — гиперпараметр, $\tau = \frac{1}{C}$ — коэффициент регуляризации.

Двухклассовая логистическая регрессия

Линейная модель классификации для двух классов $Y = \{-1, 1\}$:

$$a(x) = \text{sign}\langle w, x \rangle, \quad x, w \in \mathbb{R}^n$$

Отступ $M = \langle w, x \rangle y$

Логарифмическая функция потерь:

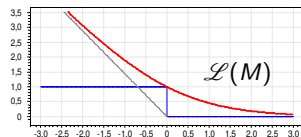
$$\mathcal{L}(M) = \log(1 + e^{-M})$$

Модель условной вероятности:

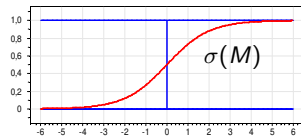
$$P(y|x, w) = \sigma(M) = \frac{1}{1 + e^{-M}},$$

$\sigma(M)$ — сигмоидная функция,

важное свойство: $\sigma(M) + \sigma(-M) = 1$



M



M

Максимизация правдоподобия (logistic loss) с регуляризацией:

$$Q(w) = \sum_{i=1}^{\ell} \log(1 + \exp(-\langle w, x_i \rangle y_i)) + \frac{\tau}{2} \|w\|^2 \rightarrow \min_w$$

Многоклассовая логистическая регрессия

Линейный классификатор при произвольном числе классов $|Y|$:

$$a(x) = \arg \max_{y \in Y} \langle w_y, x \rangle, \quad x, w_y \in \mathbb{R}^n.$$

Вероятность того, что объект x относится к классу y :

$$P(y|x, w) = \frac{\exp \langle w_y, x \rangle}{\sum_{z \in Y} \exp \langle w_z, x \rangle} = \text{SoftMax}_{y \in Y} \langle w_y, x \rangle,$$

функция SoftMax: $\mathbb{R}^Y \rightarrow \mathbb{R}^Y$ переводит произвольный вектор в нормированный вектор дискретного распределения.

Максимизация правдоподобия (log-loss) с регуляризацией:

$$L(w) = \sum_{i=1}^{\ell} \log P(y_i|x_i, w) - \frac{\tau}{2} \sum_{y \in Y} \|w_y\|^2 \rightarrow \max_w.$$

Пример. Задача кредитного скоринга

Задача кредитного скоринга:

- x_i — заёмщики
- $y_i = -1$ (bad), $+1$ (good)

Бинаризация признаков $f_j(x)$:

$$b_{jk}(x) = [f_j(x) \text{ из } k\text{-го интервала}]$$

Линейная модель классификации:

$$a(x, w) = \text{sign} \sum_{j,k} w_{jk} b_{jk}(x).$$

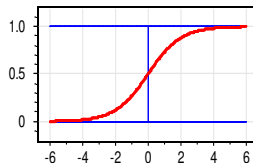
Вес признака w_{jk} равен его вкладу в общую сумму баллов (score).

признак j	интервал k	w_{jk}
Возраст	до 25	5
	25 - 40	10
	40 - 50	15
	50 и больше	10
Собственность	владелец	20
	совладелец	15
	съемщик	10
	другое	5
Работа	руководитель	15
	менеджер среднего звена	10
	служащий	5
	другое	0
Стаж	1/безработный	0
	1..3	5
	3..10	10
	10 и больше	15
Работа мужа /жены	нет/домохозяйка	0
	руководитель	10
	менеджер среднего звена	5
	служащий	1

Оценивание рисков в скоринге

Логистическая регрессия не только определяет веса w , но и оценивает *апостериорные вероятности* классов

$$P(y|x) = \frac{1}{1 + e^{-\langle w, x \rangle y}}$$



Оценка *риска* (математического ожидания) потерь объекта x :

$$R(x) = \sum_{y \in Y} D_{xy} P(y|x),$$

где D_{xy} — величина потери для объекта x с исходом y .

Оценка говорит о том, сколько мы потеряем в среднем.
Но сколько мы потеряем в худшем случае?

Методика VaR (Value at Risk)

Стохастическое моделирование: $N = 10^4$ раз

- для каждого x_i разыгрывается исход $y_i \sim P(y|x_i)$;
- вычисляется сумма потерь по портфелю $V = \sum_{i=1}^{\ell} D_{x_i y_i}$;

99%-квантиль эмпирического распределения потерь
определяет величину резервируемого капитала



Калибровка Платта (classifier with probabilistic output)

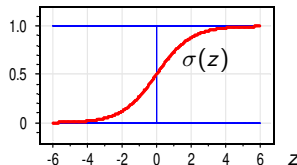
Пусть для простоты классов два, $Y = \{-1, +1\}$.

Задача. Для классификатора вида $a(x) = \text{sign } g(x, w)$ построить функцию оценки условной вероятности $P(y|x)$.

Модель условной вероятности:

$$\pi(x; \mathbf{a}, \mathbf{b}) = P(y=1|x) = \sigma(\mathbf{a}g(x, w) + \mathbf{b})$$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ — сигмоидная функция



Калибровка коэффициентов \mathbf{a}, \mathbf{b} по *контрольной* выборке методом максимума правдоподобия (снова log-loss):

$$\sum_{y_i=-1} \log(1 - \pi(x_i; \mathbf{a}, \mathbf{b})) + \sum_{y_i=+1} \log \pi(x_i; \mathbf{a}, \mathbf{b}) \rightarrow \max_{\mathbf{a}, \mathbf{b}}$$

Два подхода к обучению классификации

1 Дискриминативный (discriminative):

x — неслучайные векторы

$P(y|x, w)$ — модель классификации

Примеры: LR, GLM, SVM, RBF



2 Генеративный (generative):

$x \sim p(x|y)$ — случайные векторы

$p(x|y, w)$ — модель генерации данных

Примеры: NB, PW, FLD, RBF



Байесовские модели классификации — генеративные:

- моделируют форму классов не только вдоль границы, но и на всём пространстве, что избыточно для классификации
- требуют больше данных для обучения
- более устойчивы к шумовым выбросам

Вероятностная постановка задачи классификации

X — объекты, Y — классы, $X \times Y$ — в.п. с плотностью $p(x, y)$

Дано: $X^\ell = (x_i, y_i)_{i=1}^\ell \sim p(x, y)$ — простая выборка (i.i.d.)

Найти: $a: X \rightarrow Y$ с минимальной вероятностью ошибки

Пусть известна совместная плотность

$$p(x, y) = p(x) P(y|x) = P(y)p(x|y)$$

$P(y)$ — априорная вероятность класса y

$p(x|y)$ — плотность (функция правдоподобия) класса y

$P(y|x)$ — апостериорная вероятность класса y

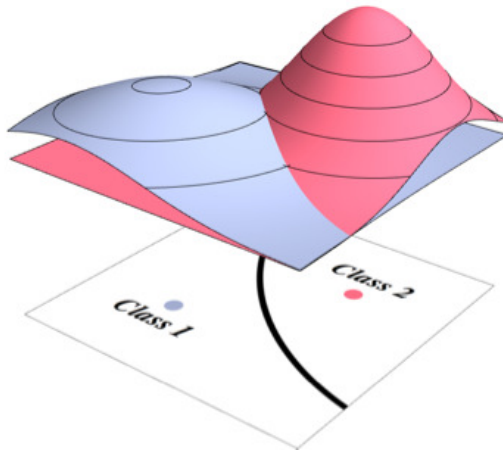
По формуле Байеса: $P(y|x) = \frac{P(y)p(x|y)}{p(x)}$

Байесовский классификатор:

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

Классификация по максимуму функции правдоподобия

Частный случай: $a(x) = \arg \max_{y \in Y} p(x|y)$ при равных $P(y)$



Оптимальный байесовский классификатор

Теорема

Пусть $P(y)$ и $p(x|y)$ известны, $\lambda_y \geq 0$ — потеря от ошибки на объекте класса $y \in Y$. Тогда минимум среднего риска

$$R(a) = \sum_{y \in Y} \lambda_y \int [a(x) \neq y] p(x, y) dx$$

достигается оптимальным байесовским классификатором

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

Замечание 1: после подстановки эмпирических оценок $\hat{P}(y)$ и $\hat{p}(x|y)$ байесовский классификатор уже не оптимален

Замечание 2: задача оценивания плотности распределения — более сложная, чем задача классификации

Наивный байесовский классификатор (Naïve Bayes)

Наивное предположение:

признаки $f_j: X \rightarrow D_j$ — независимые случайные величины с плотностями распределения, $p_j(\xi|y)$, $y \in Y$, $j = 1, \dots, n$

Тогда функции правдоподобия классов представимы в виде произведения одномерных плотностей по признакам, $x^j \equiv f_j(x)$:

$$p(x|y) = p_1(x^1|y) \cdots p_n(x^n|y), \quad x = (x^1, \dots, x^n), \quad y \in Y$$

Прологарифмировав под $\arg\max$, получим классификатор

$$a(x) = \arg \max_{y \in Y} \left(\ln \lambda_y \hat{P}(y) + \sum_{j=1}^n \ln \hat{p}_j(x^j|y) \right)$$

Восстановление n одномерных плотностей

— намного более простая задача, чем одной n -мерной

Квадратичный дискриминант (Quadratic Discriminant Analysis)

Гипотеза: каждый класс $y \in Y$ имеет n -мерную гауссовскую плотность с центром μ_y и ковариационной матрицей Σ_y :

$$p(x|y) = \mathcal{N}(x; \mu_y, \Sigma_y) = \frac{\exp\left(-\frac{1}{2}(x - \mu_y)^\top \Sigma_y^{-1}(x - \mu_y)\right)}{\sqrt{(2\pi)^n \det \Sigma_y}}$$

Теорема

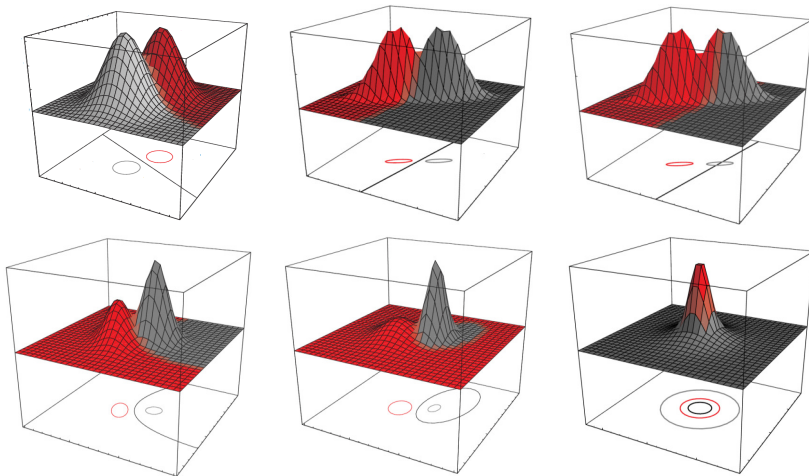
1. Разделяющая поверхность, определяемая уравнением $\lambda_y P(y) p(x|y) = \lambda_s P(s) p(x|s)$, квадратична для всех $y, s \in Y$.
2. Если $\Sigma_y = \Sigma_s$, то поверхность вырождается в линейную.

Квадратичный дискриминант — подстановочный алгоритм:

$$a(x) = \arg \max_{y \in Y} \left(\ln \lambda_y P(y) - \frac{1}{2}(x - \hat{\mu}_y)^\top \hat{\Sigma}_y^{-1}(x - \hat{\mu}_y) - \frac{1}{2} \ln \det \hat{\Sigma}_y \right)$$

Геометрический смысл квадратичного дискриминанта

Разделяющая поверхность линейна ($\Sigma_y = \Sigma_s$) или квадратична:



Линейный дискриминант Фишера (Fisher Linear Discriminant)

Проблема: для малочисленных классов возможно $\det \hat{\Sigma}_y = 0$.

Пусть ковариационные матрицы классов равны: $\Sigma_y = \Sigma$, $y \in Y$.

Оценка максимума правдоподобия для Σ :

$$\hat{\Sigma} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^T, \quad \hat{\mu}_y = \frac{\sum_i [y_i = y] x_i}{\sum_i [y_i = y]}$$

Линейный дискриминант — подстановочный алгоритм:

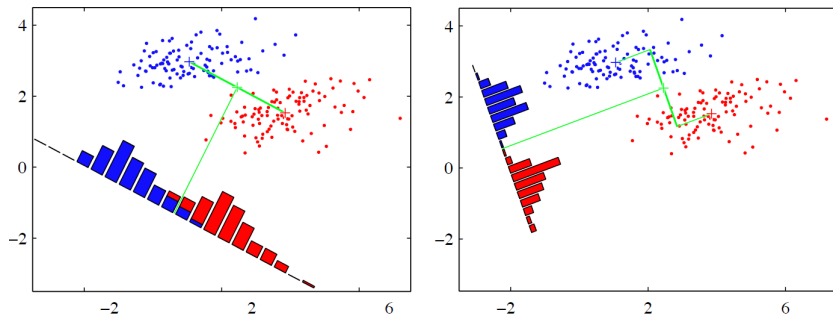
$$\begin{aligned} a(x) &= \arg \max_{y \in Y} \lambda_y \hat{P}(y) \hat{p}(x|y) = \\ &= \arg \max_{y \in Y} \underbrace{(\ln(\lambda_y \hat{P}(y)) - \frac{1}{2} \hat{\mu}_y^T \hat{\Sigma}^{-1} \hat{\mu}_y)}_{\beta_y} + x^T \underbrace{\hat{\Sigma}^{-1} \hat{\mu}_y}_{\alpha_y}; \end{aligned}$$

$$a(x) = \arg \max_{y \in Y} (x^T \alpha_y + \beta_y).$$

В случае мультиколлинеарности — обращать матрицу $\hat{\Sigma} + \tau I_n$.

Геометрическая интерпретация линейного дискриминанта

В одномерной проекции на направляющий вектор разделяющей гиперплоскости классы разделяются наилучшим образом, то есть с минимальной вероятностью ошибки:



Ось проекции перпендикулярна общей касательной эллипсоидов рассеяния

Fisher R. A. The use of multiple measurements in taxonomic problems. 1936.

Резюме в конце лекции

- Обучение вероятностных порождающих (генеративных) моделей — *методом максимума правдоподобия*
 - восстановление плотности по данным (без учителя)
 - обучение регрессии (с учителем)
 - обучение классификации (с учителем)
- Вероятностный смысл *регуляризации* — априорное распределение в пространстве параметров модели
- *Логистическая регрессия* — метод классификации, оценивающий условные вероятности классов $P(y|x)$
- Два подхода к обучению классификации:
 - *дискриминативный*: модель вероятности классов $P(y|x, w)$
 - *генеративный*: модель плотности классов $p(x|y, w)$
- Байесовские методы классификации — генеративные, основаны на оценивании плотностей классов