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-25-26
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 30 сентября 2025

Содержание

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. **Байесовская теория 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: *метод максимума правдоподобия* (ММП) выборки (Maximum Likelihood Estimate, MLE, Maximum log-Likelihood)

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

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 №1: оценка дискретного распределения

Дано: простая выборка $x_i \in X$, $|X| < \infty$, порождаемая дискретным распределением $(p_x \colon x \in X)$, $\sum_x p_x = 1$, $p_x \geqslant 0$

Найти: параметры распределения $(p_x \colon x \in X)$

Критерий: максимум (логарифма) правдоподобия выборки

$$\ln \prod_{i=1}^{\ell} p_{x_i} = \sum_{x \in X} \underbrace{\textstyle\sum_{i=1}^{\ell} [x_i = x]}_{\ell_x} \ln p_x = \sum_{x \in X} \ell_x \ln p_x \to \max_{(p_x)}$$

Выборочная оценка максимального правдоподобия

$\hat{p}_x = \frac{\ell_x}{\ell}$ — частотные оценки вероятностей $p_x = P(x_i = x)$, оценка минимума кросс-энтропии, эмпирическая гистограмма

Доказательство из условий ККТ: $\frac{\partial}{\partial p_x} \left(\sum_{x \in X} \ell_x \ln p_x + \mu \left(1 - \sum_{x \in X} p_x \right) \right) = 0$

Частный случай №2: многомерная гауссовская плотность

Дано: выборка $x_i \in \mathbb{R}^n$, порождаемая гауссовской плотностью:

$$x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(x;\mu,\Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^n \det\Sigma}}, \quad x \in \mathbb{R}^n$$

Найти параметры распределения:

$\mu \in \mathbb{R}^n$ — вектор математического ожидания, $\mu = \mathsf{E}x$

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

Критерий: максимум (логарифма) правдоподобия выборки

Выборочные оценки максимального правдоподобия:

$$\frac{\partial}{\partial\mu}\ln L(\mu,\Sigma;X^\ell) = 0 \quad\Rightarrow\quad \hat\mu = \frac{1}{\ell}\sum_{i=1}^{\ell} x_i$$

$$\frac{\partial}{\partial\Sigma}\ln L(\mu,\Sigma;X^\ell) = 0 \quad\Rightarrow\quad \hat\Sigma = \frac{1}{\ell}\sum_{i=1}^{\ell}(x_i-\hat\mu)(x_i-\hat\mu)^\mathsf{T}$$

Некоторые приёмы матричного дифференцирования

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

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

$\operatorname{diag} A$ — диагональ матрицы A, остальные элементы нули
A — квадратная $n \times n$-матрица, $\det A$ — её детерминант
u — вектор размерности n

если A произвольного вида:

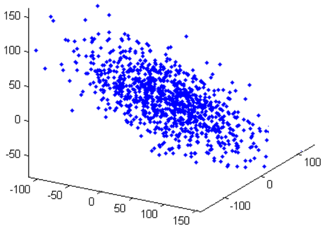
$$\frac{\partial}{\partial u} u^{\mathsf{T}} A u = A^{\mathsf{T}} u + A u$$
$$\frac{\partial}{\partial A} u^{\mathsf{T}} A u = u u^{\mathsf{T}}$$
$$\frac{\partial}{\partial A} \ln \det A = A^{-1\mathsf{T}}$$

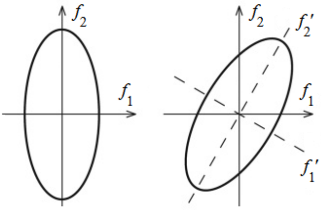
если A симметричная, $A^{\mathsf{T}} = A$:

$$\frac{\partial}{\partial u} u^{\mathsf{T}} A u = 2 A u$$
$$\frac{\partial}{\partial A} u^{\mathsf{T}} A u = 2 u u^{\mathsf{T}} - \operatorname{diag} u u^{\mathsf{T}}$$
$$\frac{\partial}{\partial A} \ln \det A = 2 A^{-1} - \operatorname{diag} A^{-1}$$

Геометрический смысл многомерной нормальной плотности

Эллипсоид рассеяния — облако точек эллиптической формы:





При $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ оси эллипсоида параллельны ортам.
В общем случае: $\Sigma = VSV^{\mathsf{T}}$ — спектральное разложение,
$V = (v_1, \ldots, v_n)$ — ортогональные собств. векторы, $V^{\mathsf{T}}V = I_n$
$S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ — собственные значения матрицы Σ

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

$x' = V^{\mathsf{T}}x$ — ортогональное преобразование поворот/отражение

Проблема мультиколлинеарности

Проблема: при $\ell < n$ матрица $\hat{\Sigma}$ вырождена, но даже при $\ell \geqslant n$ 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:

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

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

Дано: простая выборка $(x_i, y_i)_{i=1}^{\ell}$, $\;y_i = y(x_i) \in \mathbb{R}$

Найти: параметр w модели регрессии $a(x_i, w) = y(x_i) + \varepsilon_i$,
где $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ — некоррелированный гауссовский шум

Критерий: метод максимума правдоподобия (ММП)

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

$$-\ln L(\varepsilon_1, \dots, \varepsilon_\ell; w) = \mathrm{const}(w) + \frac{1}{2}\sum_{i=1}^{\ell}\frac{1}{\sigma_i^2}\bigl(a(x_i, w) - y_i\bigr)^2 \to \min_w;$$

ММП эквивалентен *методу наименьших квадратов* (МНК), если:

- модель $a(x, w)$ приближает $y(x)$ с точностью до шума ε_i
- шум ε_i гауссовский, $\mathsf{E}\varepsilon_i = 0$, некоррелированный: $\mathsf{E}\varepsilon_i\varepsilon_j = 0$
- веса объектов связаны с дисперсией шума: $\mathsf{D}\varepsilon_i = \sigma_i^2 = w_i^{-2}$

Задача классификации и принцип максимума правдоподобия

Дано: простая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$, порождаемая неизвестной плотностью $p(x, y)$ на в.п. $X \times Y$, $|Y| < \infty$

Найти: параметр w модели условной вероятности $P(y|x, w)$

$p(x, y; w) = P(y|x, w)p(x)$ — модель совместной плотности,
$p(x)$ — неизвестное и непараметризуемое распределение на X

Критерий: метод максимума правдоподобия (ММП)

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

Максимум логарифма правдоподобия (log-likelihood, log-loss):

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

Связь правдоподобия и аппроксимации эмпирического риска

Максимизация логарифма правдоподобия,
$P(y|x, w)$ — модель *условной вероятности класса*:

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

Минимизация аппроксимированного эмпирического риска,
$g(x, w)$ — модель *разделяющей поверхности*, $Y = \{\pm 1\}$:

$$Q(w) = \sum_{i=1}^{\ell} \mathscr{L}\bigl(y_i g(x_i, w)\bigr) \to \min_w;$$

Эти два принципа эквивалентны, если положить

$$-\ln P(y_i|x_i, w) = \mathscr{L}\bigl(y_i g(x_i, w)\bigr).$$

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

Вероятностный смысл регуляризации

Дано: простая выборка $(x_i, y_i)_{i=1}^{\ell} \sim p(x, y)$,
$p(w; \gamma)$ — *априорное распределение* параметров модели,
γ — вектор *гиперпараметров*

Найти: параметр w модели условной вероятности $P(y|x, w)$

Теперь случайна как выборка X^ℓ, так и модель $w \sim p(w; \gamma)$
Совместное правдоподобие данных и модели:

$$p(X^\ell, w) = p(X^\ell|w)\, p(w; \gamma)$$

Критерий *максимума апостериорной вероятности* (Maximum a Posteriori Probability, MAP):

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

Примеры: априорные распределения Гаусса и Лапласа

Линейная модель $a(x, w) = \operatorname{sign}\langle x, w\rangle$ или $\langle x, w\rangle$
Ограничения на параметры: $\mathsf{E}w_j = 0$, $\mathsf{E}w_j w_k = 0$, $\mathsf{D}w_j = C$

Распределение Гаусса и квадратичный (L_2) регуляризатор:

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

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

Распределение Лапласа и абсолютный (L_1) регуляризатор:

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

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

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

Двухклассовая (бинарная) логистическая регрессия

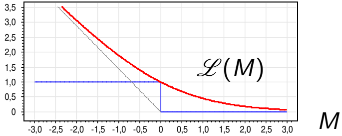
Дано: простая выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$

Найти: параметр w линейной модели $a(x, w) = \operatorname{sign}\langle x, w\rangle$

Отступ $M = \langle w, x\rangle y$

Логарифмическая функция потерь:

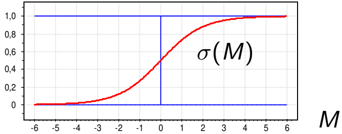
$$\mathscr{L}(M) = \ln\bigl(1 + e^{-M}\bigr)$$



Модель условной вероятности:

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

$\sigma(M)$ — сигмоидная функция,



Критерий: максимум регуляризованного log правдоподобия:

$$Q(w) = \sum_{i=1}^{\ell} \ln\bigl(1 + \exp(-\langle w, x_i\rangle y_i)\bigr) + \frac{\tau}{2}\|w\|^2 \;\to\; \min_w$$

Многоклассовая логистическая регрессия

Дано: простая выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in Y$, $2 \leqslant |Y| < \infty$

Найти: линейную модель классификации

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

и вероятность того, что объект x относится к классу y:

$$P(y|x, w) = \frac{\exp\langle w_y, x \rangle}{\sum\limits_{z \in Y} \exp\langle w_z, x \rangle} = \underset{y \in Y}{\mathrm{SoftMax}} \langle w_y, x \rangle,$$

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

Критерий: максимум регуляризованного log правдоподобия:

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

Скоринг — линейная вероятностная модель принятия решений

Пример. Кредитный скоринг:

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

Бинаризация признаков $f_j(x)$:

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

Линейная модель классификации:

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

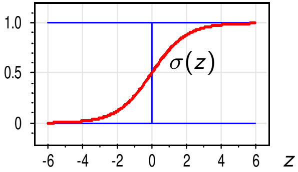
Вес признака w_{jk} равен его вкладу в общую сумму баллов (score).

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:

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



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

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

где D_{xy} — величина потери для объекта x с исходом y, причём если $y = -1$ (bad), то $D_{xy} > 0$; если $y = +1$ (good), то $D_{xy} < 0$

Оценка $R(x)$ говорит о том, сколько мы потеряем в среднем. Но сколько мы рискуем потерять в 1% худших случаев?

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)

Дано: простая выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$;
ранее построенная модель классификации $a(x) = \operatorname{sign} g(x, w)$

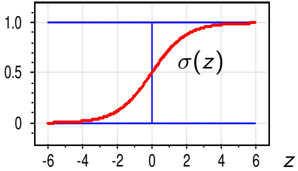
Найти: вероятностную модель классификации $\mathsf{P}(y|x)$

Модель условной вероятности:

$\pi(x; a, b) = \mathsf{P}(y{=}1|x) = \sigma\bigl(ag(x, w) + b\bigr)$

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

важное свойство: $\sigma(-z) = 1 - \sigma(z)$

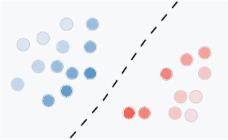


Критерий: максимум log-правдоподобия для калибровки коэффициентов a, b по *контрольной* выборке (hold-out):

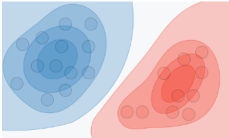
$$\sum_{y_i=-1} \ln\bigl(1 - \pi(x_i; a, b)\bigr) + \sum_{y_i=+1} \ln \pi(x_i; a, b) \to \max_{a,b}$$

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
- как правило, более устойчивы к шумовым выбросам

Вероятностные генеративные модели классификации

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

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

Найти: модель классификации $a\colon X \to Y$

Критерий: минимизировать вероятность ошибки $\mathsf{P}[a(x) \neq y]$

Пусть известна совместная плотность

$$p(x, y) = p(x)\,\mathsf{P}(y|x) = \mathsf{P}(y)p(x|y)$$

$\mathsf{P}(y)$ — *априорная вероятность* класса y
$p(x|y)$ — *функция правдоподобия* класса y
$\mathsf{P}(y|x)$ — *апостериорная вероятность* класса y

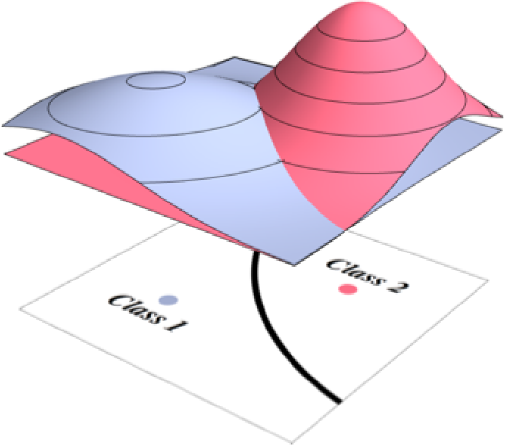
По формуле Байеса: $\mathsf{P}(y|x) = \frac{P(y)p(x|y)}{p(x)}$

Байесовский классификатор:

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

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 $\mathsf{P}(y)$ и $p(x|y)$ известны, $\lambda_y \geqslant 0$ — *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

Байесовский классификатор и максимизация правдоподобия

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

Найти: параметры $\{P_y, w_y\}$ модели $p(x, y) = P_y p(x|y; w_y)$

Критерий: метод максимума правдоподобия (ММП)

$$\sum_{i=1}^{\ell} \ln p(x_i, y_i) = \sum_{i=1}^{\ell} \ln P_{y_i} + \sum_{y\in Y} \sum_{i\in I_y} \ln p(x_i|y; w_y) \quad \to \max_{\{P_y, w_y\}}$$

Декомпозиция на $|Y| + 1$ независимо решаемых подзадач:

$$\sum_{i=1}^{\ell} \ln P_{y_i} \to \max: \left\{\textstyle\sum_y P_y = 1;\ P_y \geqslant 0\right\} \quad \Rightarrow \quad P_y = \hat{P}(y) = \frac{|I_y|}{\ell}$$

$$\sum_{i\in I_y} \ln p(x_i|y; w_y) \to \max_{w_y} \quad \text{— MLE оценка плотности класса } y$$

где $I_y = \{i\colon y_i = y\}$ — все объекты выборки класса y

Наивный байесовский классификатор (Naïve Bayes)

Наивное предположение:
признаки $f_j\colon X \to D_j$ — независимые случайные величины с плотностями распределения, $p_j(\xi|y)$, $y \in Y$, $j = 1, \ldots, n$

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

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

Прологарифмировав под argmax, 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:

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

Теорема

1. Разделяющая поверхность, определяемая уравнением $\lambda_y P(y) p(x|y) = \lambda_s P(s) p(x|s)$, квадратична для всех $y, s \in Y$.

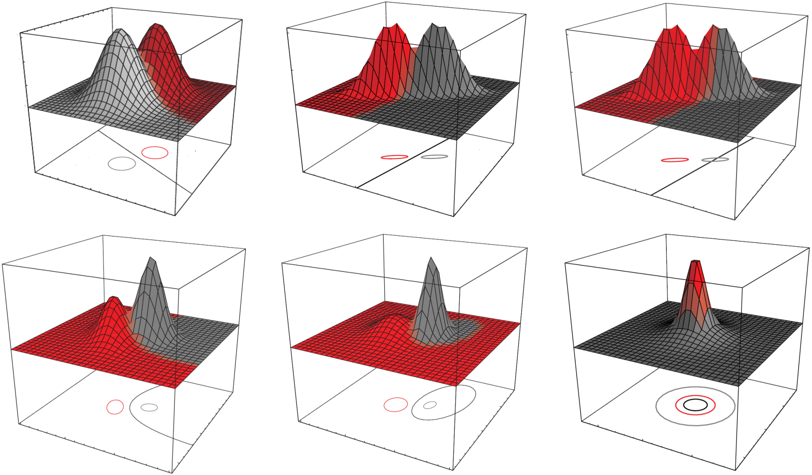
2. Если $\Sigma_y = \Sigma_s$, то поверхность вырождается в линейную.

Квадратичный дискриминант — подстановочный алгоритм:

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

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:

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

Линейный дискриминант — подстановочный алгоритм:

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

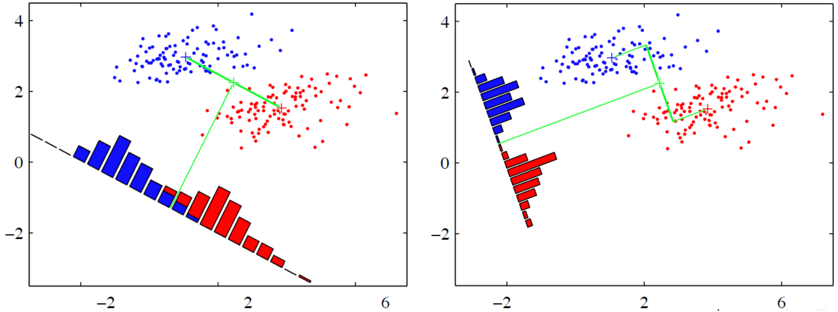
$$= \arg\max_{y \in Y} \Bigl(\underbrace{\ln(\lambda_y \hat{P}(y)) - \tfrac{1}{2} \hat{\mu}_y^{\mathsf{T}} \hat{\Sigma}^{-1} \hat{\mu}_y}_{\beta_y} + x^{\mathsf{T}} \underbrace{\hat{\Sigma}^{-1} \hat{\mu}_y}_{\alpha_y} \Bigr);$$

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

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

- *Метод максимума правдоподобия* — основной инструмент обучения вероятностных моделей для различных задач:
  — 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 апостериорные 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, сводятся к оцениванию плотностей классов