• Введение в машинное обучение •

Вероятностные модели машинного обучения

Воронцов Константин Вячеславович
k.v.vorontsov@phystech.edu
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Введение в машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ.ФПМИ.ИС.ИАД • 2 апреля 2016

Основные научные школы машинного обучения

1. *символизм* – поиск логических закономерностей
   - Decision Tree, Rule Induction
2. *коннекционизм* – обучаемые нейронные сети
   - BackPropagation, Deep Belief Nets, Deep Learning
     CNN, ResNet, LSTM, GRU, Attention, Transformer
3. *эволюционизм* – саморазвитие сложных моделей
   - Genetic Algorithms, Genetic Programming, Symbolic Regression
4. *байесионизм* и вероятностно-статистические методы
   - MLE, EM, GLM, LR, OBC, Naive Bayes, QD, LDF
     Bayesian Networks, Bayesian Learning, Graphical Models
5. *аналогизм* – «близким объектам близкие ответы»
   - kNN, RBF, SVM, KDE, Kernel Smoothing

⊕ *композиционизм* – кооперация моделей
   - Weighted Voting, Boosting, Bagging, Stacking,
     Random Forest, Яндекс.CatBoost

Педро Домингос. Верховный алгоритм. 2016. 336 с.
Pedro Domingos. The master algorithm. 2015. 329 p.

Содержание

Задача оценивания плотности — обучение без учителя

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

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):

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

Аналитическое решение: необходимое условие экстремума

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

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

Частный случай: оценка многомерной гауссовской плотности

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

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

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

Выборочные оценки максимального правдоподобия:

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

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

Частный случай: оценка дискретного распределения

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

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

Критерий: максимум (логарифма) правдоподобия выборки

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

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

Задача разделения смеси распределений

Дано: выборка $\{x_1, \ldots, x_\ell\}$, порождаемая i.i.d. из $p(x)$

Найти: модель вероятностной смеси k распределений:

$$p(x) = \sum_{j=1}^{k} w_j \varphi(x, \theta_j), \qquad \sum_{j=1}^{k} w_j = 1, \qquad w_j \geqslant 0,$$

описывающую двухуровневый процесс порождения данных:

1. $j \sim P(j) \equiv w_j$ — дискретное *априорное распределение*
2. $x \sim p(x|j) \equiv \varphi(x, \theta_j)$ — плотность j-й компоненты

Критерий максимума log-правдоподобия:

$$\begin{cases} Q(w, \theta) = \ln \prod_{i=1}^{\ell} p(x_i) = \sum_{i=1}^{\ell} \ln \sum_{j=1}^{k} w_j \varphi(x_i, \theta_j) \to \max\limits_{w,\theta} \\ \sum_{j=1}^{k} w_j = 1, \quad w_j \geqslant 0 \end{cases}$$

EM-алгоритм для разделения смеси распределений

Теорема (необходимые условия экстремума)

Точка $(w_j, \theta_j)_{j=1}^k$ локального экстремума $Q(w, \theta)$ удовлетворяет системе уравнений относительно параметров модели w_j, θ_j и вспомогательных переменных g_{ij}:

$$\text{E-шаг:} \quad g_{ij} = \frac{w_j \varphi(x_i, \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i, \theta_s)}, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, k;$$

$$\text{M-шаг:} \quad \theta_j = \arg\max_\theta \sum_{i=1}^\ell g_{ij} \ln \varphi(x_i, \theta), \quad j = 1, \ldots, k;$$

$$w_j = \frac{1}{\ell} \sum_{i=1}^\ell g_{ij}, \quad j = 1, \ldots, k.$$

EM-алгоритм — метод простых итераций для решения системы

Доказательство — через условия Каруша–Куна–Таккера

Лагранжиан оптимизационной задачи $Q(w,\theta) \to \max$:

$$\mathscr{L}(w,\theta) = \sum_{i=1}^{\ell} \ln \Bigg(\underbrace{\sum_{j=1}^{k} w_j \varphi(x_i,\theta_j)}_{p(x_i)} \Bigg) - \lambda \Bigg(\sum_{j=1}^{k} w_j - 1 \Bigg)$$

Приравниваем нулю производные:

$$\frac{\partial \mathscr{L}}{\partial w_j} = 0 \quad \Rightarrow \quad \sum_{i=1}^{\ell} \underbrace{\frac{w_j \varphi(x_i,\theta_j)}{p(x_i)}}_{g_{ij}} = \lambda w_j; \quad \lambda = \ell; \quad w_j = \frac{1}{\ell} \sum_{i=1}^{\ell} g_{ij}$$

$$\frac{\partial \mathscr{L}}{\partial \theta_j} = \sum_{i=1}^{\ell} \underbrace{\frac{w_j \varphi(x_i,\theta_j)}{p(x_i)}}_{g_{ij}} \frac{\frac{\partial}{\partial \theta_j} \varphi(x_i,\theta_j)}{\varphi(x_i,\theta_j)} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{\ell} g_{ij} \ln \varphi(x_i,\theta_j) = 0$$

■

Вероятностная интерпретация шагов EM-алгоритма

E-шаг — это формула Байеса:

$$g_{ij} = \mathrm{P}(j|x_i) = \frac{P(j)p(x_i|j)}{p(x_i)} = \frac{w_j\varphi(x_i,\theta_j)}{p(x_i)} = \frac{w_j\varphi(x_i,\theta_j)}{\sum_{s=1}^{k} w_s\varphi(x_i,\theta_s)}$$

Нормировка условных вероятностей: $\sum_{j=1}^{k} g_{ij} = 1$

M-шаг — это максимизация взвешенного log-правдоподобия, с весами объектов g_{ij} для j-й компоненты смеси:

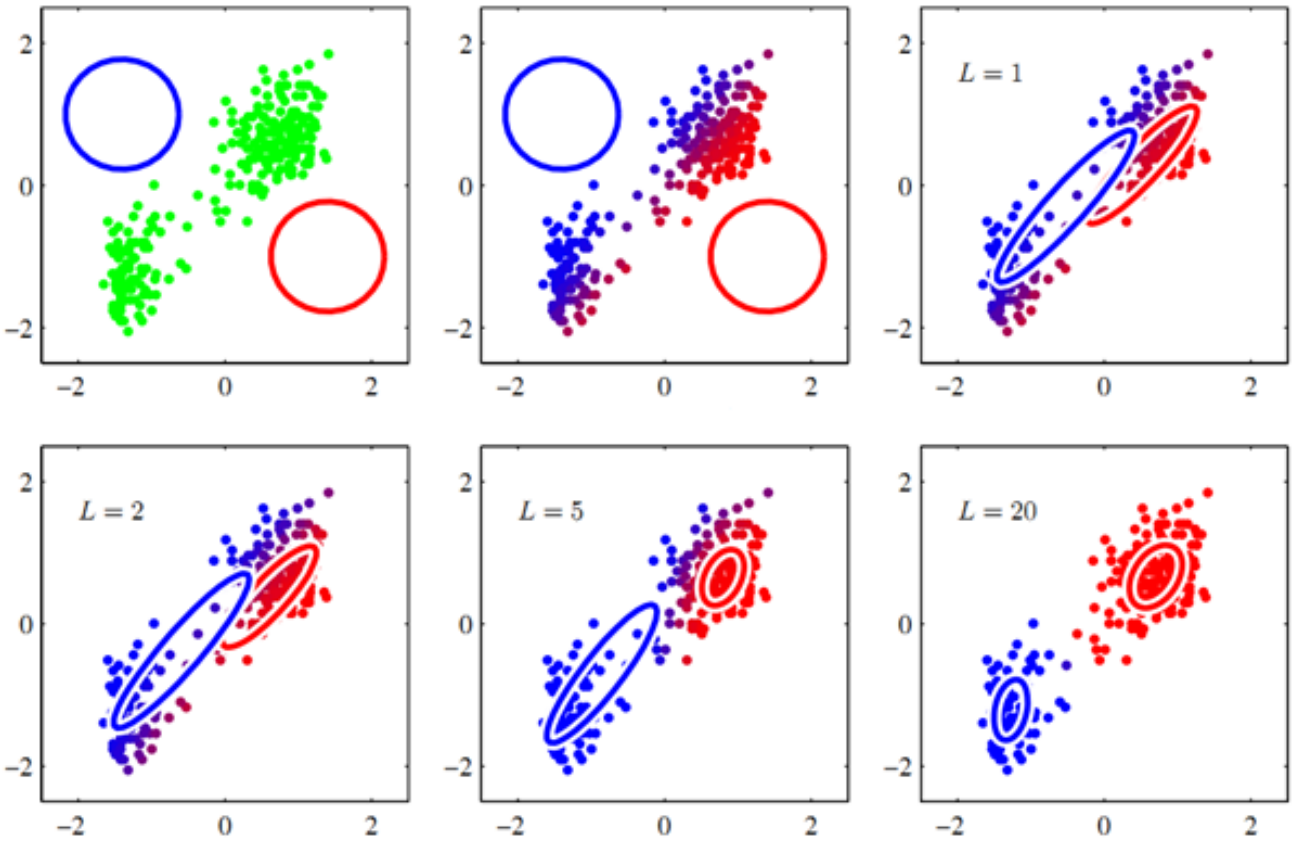
$$\theta_j = \arg\max_{\theta} \sum_{i=1}^{\ell} g_{ij} \ln \varphi(x_i,\theta),$$

вес компоненты определяется как средний вес её объектов:

$$w_j = \frac{1}{\ell} \sum_{i=1}^{\ell} g_{ij}$$

Пример: разделение гауссовской смеси

Две гауссовские компоненты $k = 2$ в пространстве $X = \mathbb{R}^2$.
Расположение компонент в зависимости от номера итерации L:



Вероятностная постановка задачи регрессии

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

Найти: параметр w модели регрессии с гауссовским шумом:

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2),\ \mu_i = a(x_i, w) = \mathsf{E}y_i, \quad i = 1, \ldots, \ell.$$

Эквивалентная запись: $y_i = a(x_i, w) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

Критерий максимума правдоподобия эквивалентен МНК:

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

$$-\ln p(\varepsilon_1, \ldots, \varepsilon_\ell | w) = \text{const} + \frac{1}{2}\sum_{i=1}^{\ell} \frac{1}{\sigma_i^2}\big(a(x_i, w) - y_i\big)^2 \to \min_w$$

Что использовать вместо метода наименьших квадратов, если y_i не гауссовские, в частности, если y_i дискретнозначные?

Обобщённая линейная модель (Generalized Linear Model, GLM)

Дано: выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$
Найти: параметр w *обобщённой линейной модели* (GLM):

$$y_i \sim \text{Exp}(\theta_i, \varphi_i), \qquad \theta_i = g(\text{E}y_i) = a(x_i, w) = \langle x_i, w \rangle,$$

вместо предположения о гауссовости y_i, теперь вводится Exp — *экспоненциальное семейство распределений* (exponential family) с параметром θ_i, параметром масштаба φ_i (scale) и параметрами-функциями $c(\theta)$, $h(y, \varphi)$:

$$p(y_i|x_i) = \exp\left(\frac{y_i\theta_i - c(\theta_i)}{\varphi_i} + h(y_i, \varphi_i)\right)$$

Критерий максимума правдоподобия:

$$Q(w) = \ln \prod_{i=1}^{\ell} p(y_i|x_i) = \sum_{i=1}^{\ell} \frac{y_i\langle x_i, w\rangle - c\big(\langle x_i, w\rangle\big)}{\varphi_i} \to \max_{w,\{\varphi_i\}}$$

Экспоненциальное семейство распределений

Exp — *экспоненциальное семейство распределений*
с параметрами θ_i, φ_i и параметрами-функциями $c(\theta)$, $h(y, \varphi)$:

$$p(y_i|\theta_i, \varphi_i) = \exp\left(\frac{y_i\theta_i - c(\theta_i)}{\varphi_i} + h(y_i, \varphi_i)\right)$$

Свойства экспоненциальных распределений

Математическое ожидание и дисперсия с.в. $y_i \sim \text{Exp}(\theta_i, \varphi_i)$:

$$\mu_i = \text{E}y_i = c'(\theta_i) \quad \Rightarrow \quad \theta_i = [c']^{-1}(\mu_i) = g(\text{E}y_i)$$
$$\text{D}y_i = \varphi_i c''(\theta_i)$$

$g(\mu) = [c']^{-1}(\mu)$ — монотонная *функция связи* (link function)

Нормальная линейная модель — частный случай GLM:

$$y_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad \mu_i = \langle x_i, w\rangle = \text{E}y_i \qquad g(\mu_i) = \mu_i$$

Примеры распределений из экспоненциального семейства

Нормальное (гауссовское) распределение, $y_i \in \mathbb{R}$:

$$p(y_i|\mu_i,\sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\Big(-\frac{1}{2\sigma_i^2}(y_i-\mu_i)^2\Big) =$$

$$= \exp\left(\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma_i^2} - \frac{y_i^2}{2\sigma_i^2} - \frac{1}{2}\ln(2\pi\sigma_i^2)\right);$$

$$\theta_i = g(\mu_i) = \mu_i, \qquad c(\theta_i) = \tfrac{1}{2}\mu_i^2 = \tfrac{1}{2}\theta_i^2, \qquad \varphi_i = \sigma_i^2.$$

Распределение Бернулли, $y_i \in \{0,1\}$:

$$p(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \exp\left(y_i \ln\tfrac{\mu_i}{1-\mu_i} + \ln(1-\mu_i)\right);$$

$$\theta_i = g(\mu_i) = \ln\tfrac{\mu_i}{1-\mu_i}, \qquad c(\theta_i) = -\ln(1-\mu_i) = \ln(1+e^{\theta_i}).$$

Примеры распределений из экспоненциального семейства

Биномиальное распределение, $y_i \in \{0, 1, \ldots, n_i\}$:

$$p(y_i|\mu_i, n_i) = C_{n_i}^{y_i} \left(\frac{\mu_i}{n_i}\right)^{y_i} \left(1 - \frac{\mu_i}{n_i}\right)^{n_i - y_i} =$$

$$= \exp\left(y_i \ln \frac{\mu_i}{n_i - \mu_i} + n_i \ln(n_i - \mu_i) + \ln C_{n_i}^{y_i} - n_i \ln n_i\right);$$

$$\theta_i = g(\mu_i) = \ln \frac{\mu_i}{n_i - \mu_i}, \qquad c(\theta_i) = -n_i \ln(n_i - \mu_i) = n_i \ln \frac{1 + e^{\theta_i}}{n_i}.$$

Пуассоновское распределение, $y_i \in \{0, 1, 2, \ldots\}$:

$$p(y_i|\mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \exp\left(\frac{y_i \ln(\mu_i) - \mu_i}{1} - \ln y_i!\right);$$

$$\theta_i = g(\mu_i) = \ln(\mu_i), \qquad c(\theta_i) = \mu_i = e^{\theta_i}, \qquad \varphi_i = 1.$$

Примеры распределений из экспоненциального семейства

- нормальное (гауссовское)
- распределение Пуассона
- биномиальное и мультиномиальное
- геометрическое
- χ^2-распределение
- бета-распределение
- гамма-распределение
- распределение Дирихле
- распределение Лапласа с фиксированным матожиданием

Контр-примеры не экспоненциальных распределений:

- t-распределение Стьюдента, Коши, гипергеометрическое

Двухклассовая логистическая регрессия

Распределение Бернулли, $y_i \in \{0,1\}$: $p(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}$

$\theta_i = g(\mu_i) = \ln \frac{\mu_i}{1-\mu_i}$ $\qquad$ $\mathsf{E}y_i = \mu_i = g^{-1}(\theta_i) = \frac{1}{1+\exp(-\theta_i)} \equiv \sigma(\theta_i)$

Дано: выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \{0,1\} \sim p(y_i|\mu_i)$

Найти: вероятностную модель $p(y|x) = \mathsf{E}(y|x) = \sigma\big(\langle x, w\rangle\big)$

Критерий: максимум log-правдоподобия (log-loss)

$$Q(w) = \sum_{i=1}^{\ell} \ln p(y_i|\mu_i) = \sum_{i=1}^{\ell} y_i \ln \mu_i + (1-y_i)\ln(1-\mu_i) \to \max_w$$

Удобная перекодировка: $y_i \in \{0,1\} \;\to\; \tilde{y}_i = 2y_i - 1 \in \{-1,1\}$

$$-\sum_{i=1}^{\ell} \ln p(\tilde{y}_i|x_i) = \sum_{i=1}^{\ell} \ln\big(1+\exp(-\underbrace{\langle w, x_i\rangle \tilde{y}_i}_{\text{margin}})\big) \;\to\; \min_w$$

Логистическая регрессия как частный случай GLM

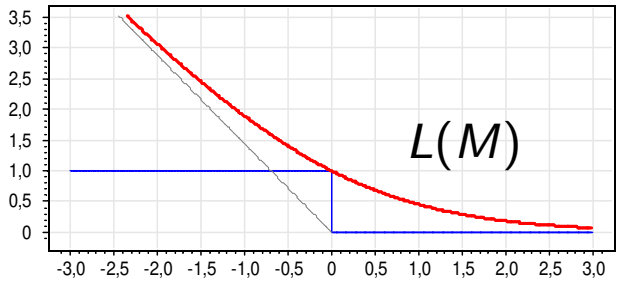
Всего лишь из двух предположений:

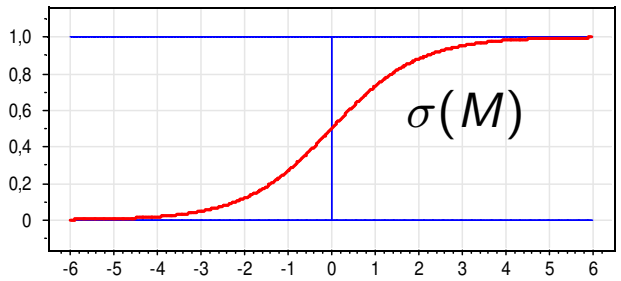
- y_i — бернуллиевские случайные величины с $\mathsf{E}y_i = \mu_i$
- параметр связан с линейной моделью: $\theta_i = g(\mu_i) = \langle x_i, w \rangle$

следуют важнейшие свойства логистической регрессии:

- логарифмическая функция потерь $\ln\bigl(1 + \exp(-\langle x_i, w\rangle \tilde{y}_i)\bigr)$;
- сигмоидная функция связи $\mathsf{P}(y_i|x_i) = \sigma(\langle x_i, w\rangle \tilde{y}_i)$;
- связь линейной модели с *отношением шансов* (odds ratio):

$$\langle x_i, w\rangle = \ln \frac{\mu_i}{1-\mu_i} = \ln \frac{\mathsf{P}(y_i=1|x_i)}{\mathsf{P}(y_i=0|x_i)}$$





Многоклассовая логистическая регрессия

Дано: 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

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

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

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

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

Критерий: максимум log-правдоподобия (log-loss)

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

Калибровка Платта (classifier with probabilistic output)

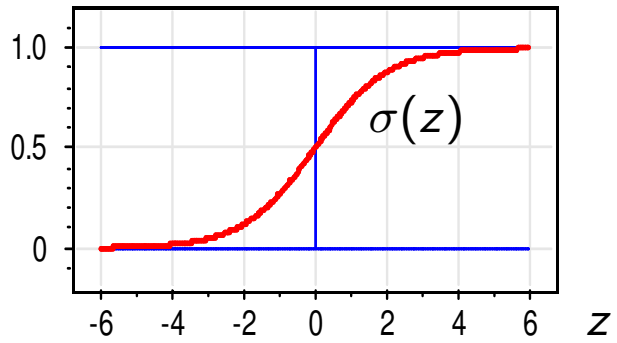
Дано: выборка $(x_i, y_i)_{i=1}^{\ell}$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$;
ранее построенная модель классификации $a(x) = \operatorname{sign} g(x, w)$

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

Модель условной вероятности:

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

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



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

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

Скоринг — линейная вероятностная модель принятия решений

Пример. Кредитный скоринг:

- x_i — заёмщики
- $y_i = -1\,(\text{bad}), +1\,(\text{good})$

Бинаризация признаков $f_j(x)$:

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

Линейная модель классификации:

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

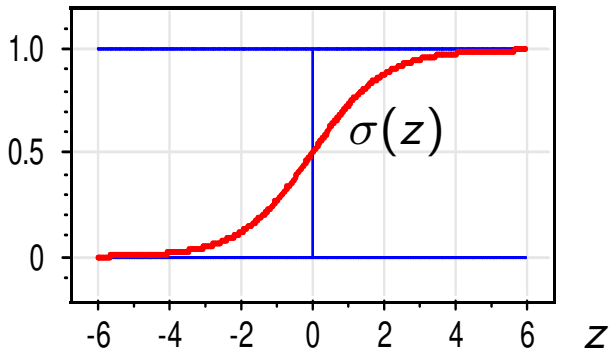
Вес признака w_{jk} равен его вкладу в общую сумму баллов (score).

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:

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



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

Вероятностные дискриминативные модели классификации

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

Найти: модель плотности $p(x, y; w) = P(y|x, w)p(x)$, где $P(y|x, w)$ — модель условной вероятности класса с параметром w
$p(x)$ — неизвестное и непараметризуемое распределение на X

Критерий: максимум правдоподобия, т.е. max плотности совместного распределения всей выборки данных:

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

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

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

Связь правдоподобия и аппроксимации эмпирического риска

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

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

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

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

Эти два принципа эквивалентны, если положить

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

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

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:

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

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

$$Q_{\text{MAP}}(w) = \ln p(X^\ell, w) = \underbrace{\sum_{i=1}^{\ell} \ln P(y_i|x_i,w)}_{Q_{\text{MLE}}(w)} + \underbrace{\ln p(w;\gamma)}_{\text{регуляризатор}} \to \max_w$$

Примеры: априорные распределения Гаусса и Лапласа

Пусть веса w_j независимы, $\mathsf{E}w_j = 0$, $\mathsf{D}w_j = C$.

Распределение Гаусса и квадратичный (L_2) регуляризатор:

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

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

Распределение Лапласа и абсолютный (L_1) регуляризатор:

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

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

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

Вероятностные генеративные модели 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: модель классификации $a\colon X \to Y$

Критерий: минимальная вероятность ошибки

Пусть известна совместная плотность

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

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

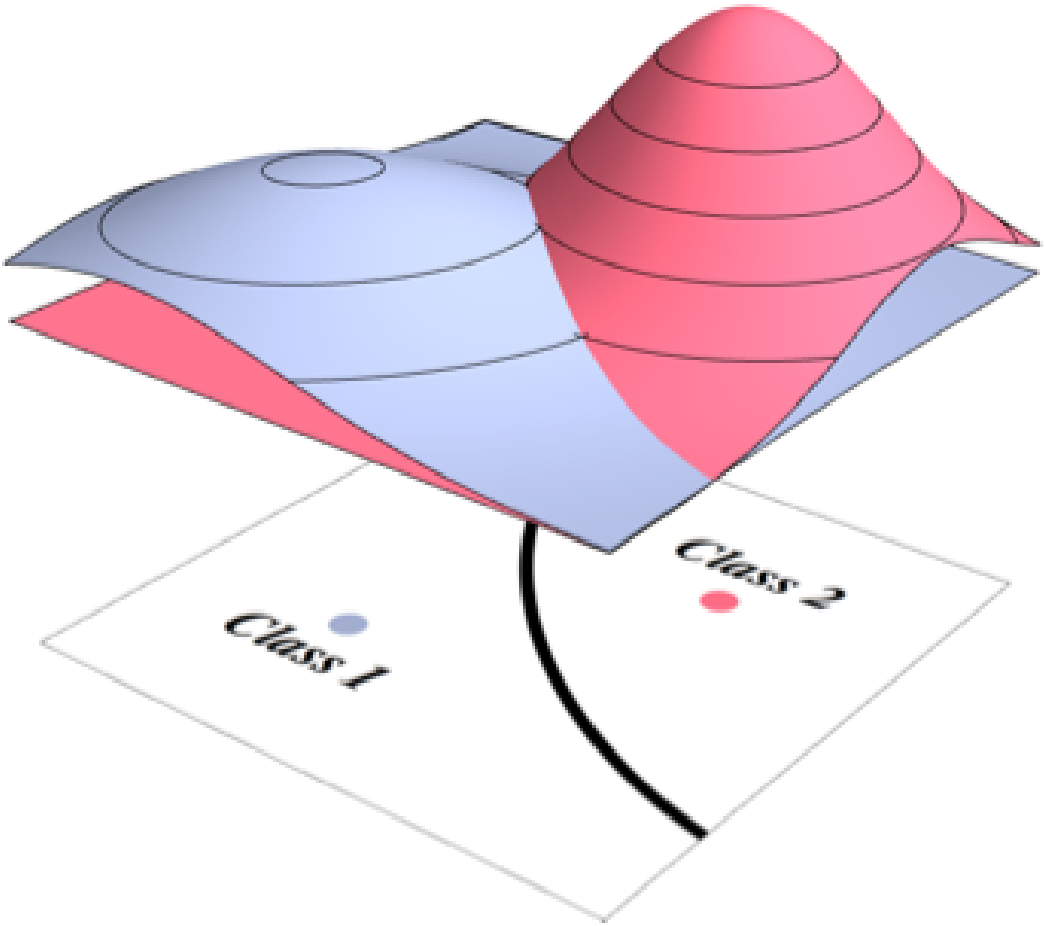
По формуле Байеса: $\mathrm{P}(y|x) = \dfrac{P(y)p(x|y)}{p(x)}$

Байесовский классификатор:

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

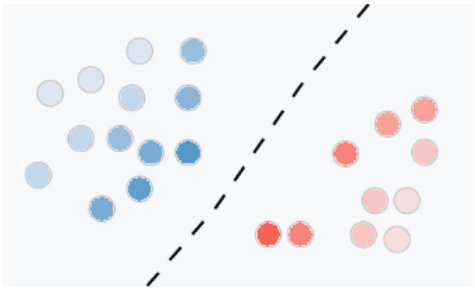
Классификация по максимуму функции правдоподобия

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

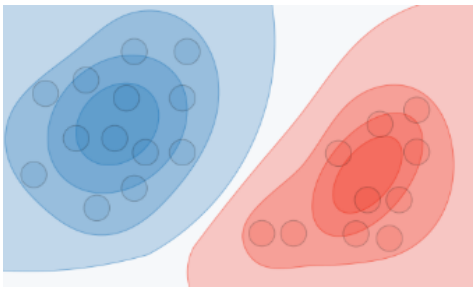


Два подхода к обучению классификации

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
   $p(x|y,\theta)$ — 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

Оптимальность байесовского классификатора

Теорема (Optimal Bayesian Classifier, OBC)

Пусть $P(y)$ 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

Наивный байесовский классификатор (Naïve Bayes)

Наивное предположение:
признаки $f_j: X \to D_j$ — 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

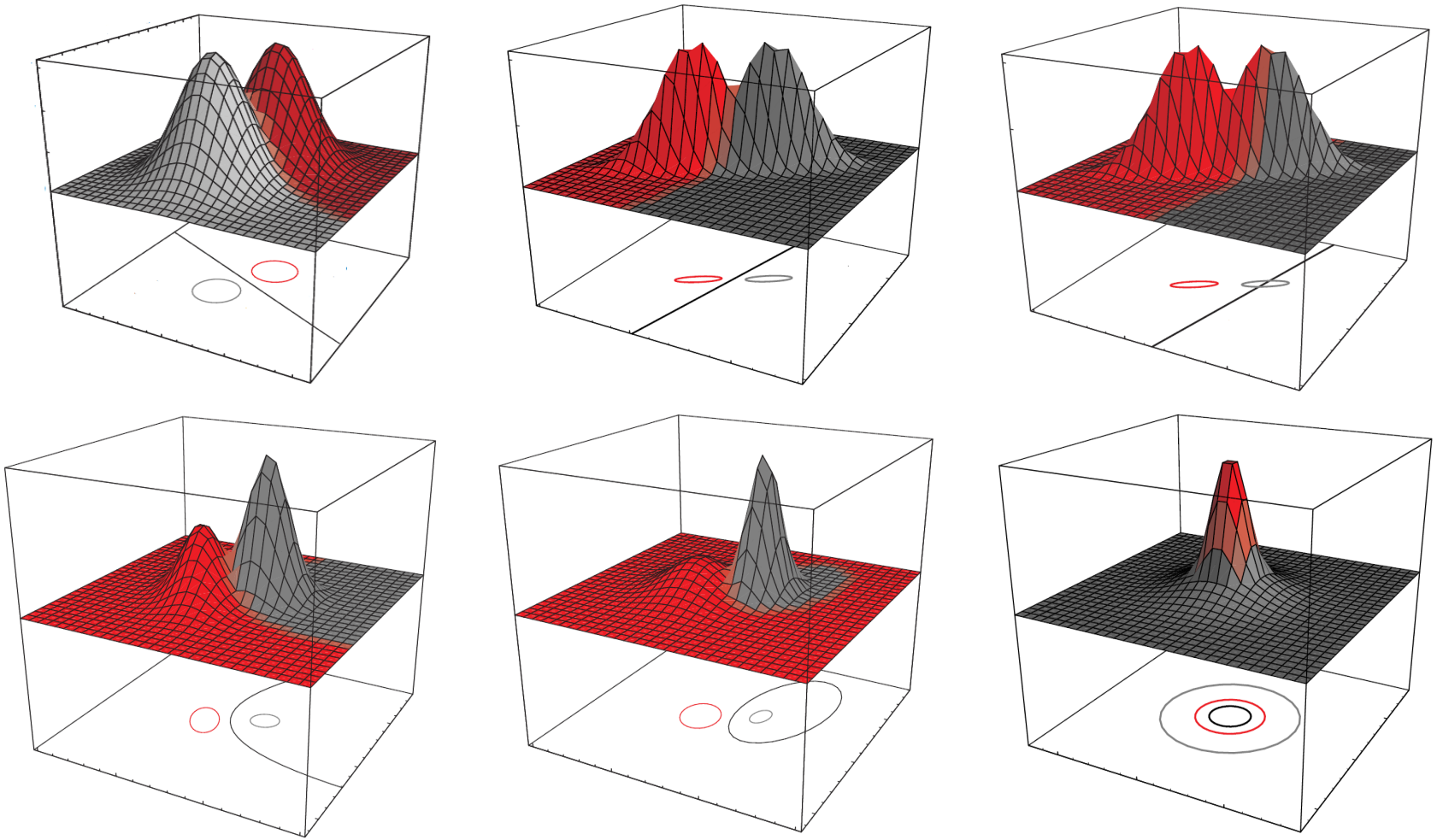
1. Разделяющая поверхность, определяемая уравнением $\lambda_y P(y)p(x|y) = \lambda_s P(s)p(x|s)$, квадратична для всех $y, s \in Y$.

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

Квадратичный дискриминант — подстановочный алгоритм:

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

Геометрический смысл квадратичного дискриминанта



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:

$$\hat{\Sigma} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^{\mathsf{T}}$$

Линейный дискриминант — подстановочный алгоритм:

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

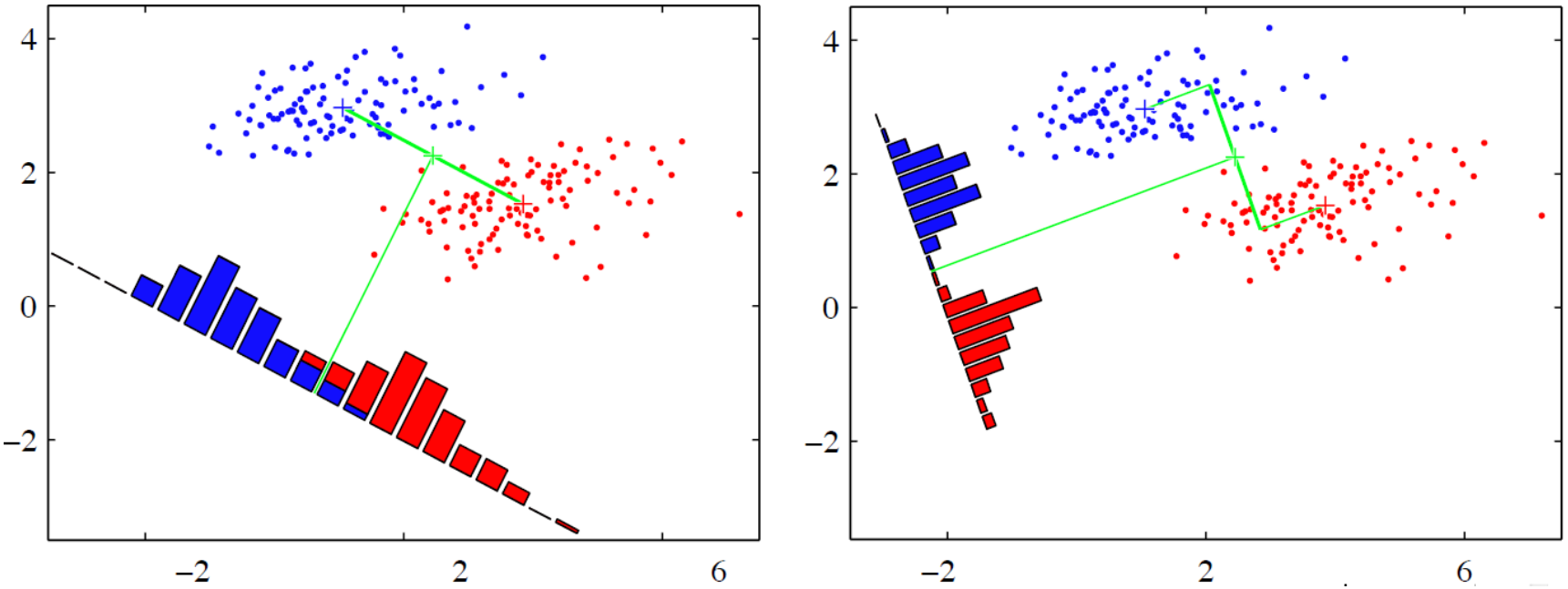
$$= \arg\max_{y \in Y} \Big(\underbrace{\ln(\lambda_y \hat{P}(y)) - \tfrac{1}{2} \hat{\mu}_y^{\mathsf{T}} \hat{\Sigma}^{-1} \hat{\mu}_y}_{\beta_y} + x^{\mathsf{T}} \underbrace{\hat{\Sigma}^{-1} \hat{\mu}_y}_{w_y} \Big);$$

$$a(x) = \arg\max_{y \in Y} \left(x^{\mathsf{T}} w_y + \beta_y \right).$$

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.

Гауссовская смесь с диагональными матрицами ковариации

Гауссовская смесь GMM — Gaussian Mixture Model

Допущения:

1. Функции правдоподобия классов $p(x|y)$ представимы в виде смесей k_y компонент, для каждого класса $y \in Y$
2. Компоненты $j = 1, \ldots, k_y$ имеют n-мерные гауссовские плотности с некоррелированными признаками: $\mu_{yj} = (\mu_{yj1}, \ldots, \mu_{yjn})$, $\Sigma_{yj} = \text{diag}(\sigma^2_{yj1}, \ldots, \sigma^2_{yjn})$:

$$p(x|y) = \sum_{j=1}^{k_y} w_{yj} p_{yj}(x), \quad p_{yj}(x) = \mathcal{N}(x; \mu_{yj}, \Sigma_{yj})$$

$$\sum_{j=1}^{k_y} w_{yj} = 1, \quad w_{yj} \geqslant 0$$

EM-алгоритм. Эмпирические оценки средних и дисперсий

Числовые признаки: $f_d : X \to \mathbb{R}$, $d = 1, \ldots, n$.

E-шаг: для всех $y \in Y$, $j = 1, \ldots, k_y$, $d = 1, \ldots, n$:

$$g_{yij} = \frac{w_{yj}\mathcal{N}(x_i; \mu_{yj}, \Sigma_{yj})}{p(x_i|y)} \equiv \mathrm{P}(j|x_i, y_i = y)$$

M-шаг: для всех $y \in Y$, $j = 1, \ldots, k_y$, $d = 1, \ldots, n$

$$w_{yj} = \frac{1}{\ell_y} \sum_{i:\, y_i = y} g_{yij}$$

$$\hat{\mu}_{yjd} = \frac{1}{\ell_y w_{yj}} \sum_{i:\, y_i = y} g_{yij} f_d(x_i)$$

$$\hat{\sigma}^2_{yjd} = \frac{1}{\ell_y w_{yj}} \sum_{i:\, y_i = y} g_{yij} \big(f_d(x_i) - \hat{\mu}_{yjd}\big)^2$$

Замечание: компоненты «наивны», но смесь не «наивна»

Байесовский классификатор

Подставим гауссовскую смесь в байесовский классификатор:

$$a(x) = \arg\max_{y \in Y} \underbrace{\lambda_y P_y \sum_{j=1}^{k_y} w_{yj} \underbrace{\mathcal{N}_{yj} \exp\left(-\tfrac{1}{2}\rho^2_{yj}(x, \mu_{yj})\right)}_{p_{yj}(x)}}_{\Gamma_y(x)}$$

$\mathcal{N}_{yj} = (2\pi)^{-\frac{n}{2}}(\sigma_{yj1} \cdots \sigma_{yjn})^{-1}$ — нормировочные множители;
$\rho_{yj}(x, \mu_{yj})$ — *взвешенная евклидова метрика* в $X = \mathbb{R}^n$:

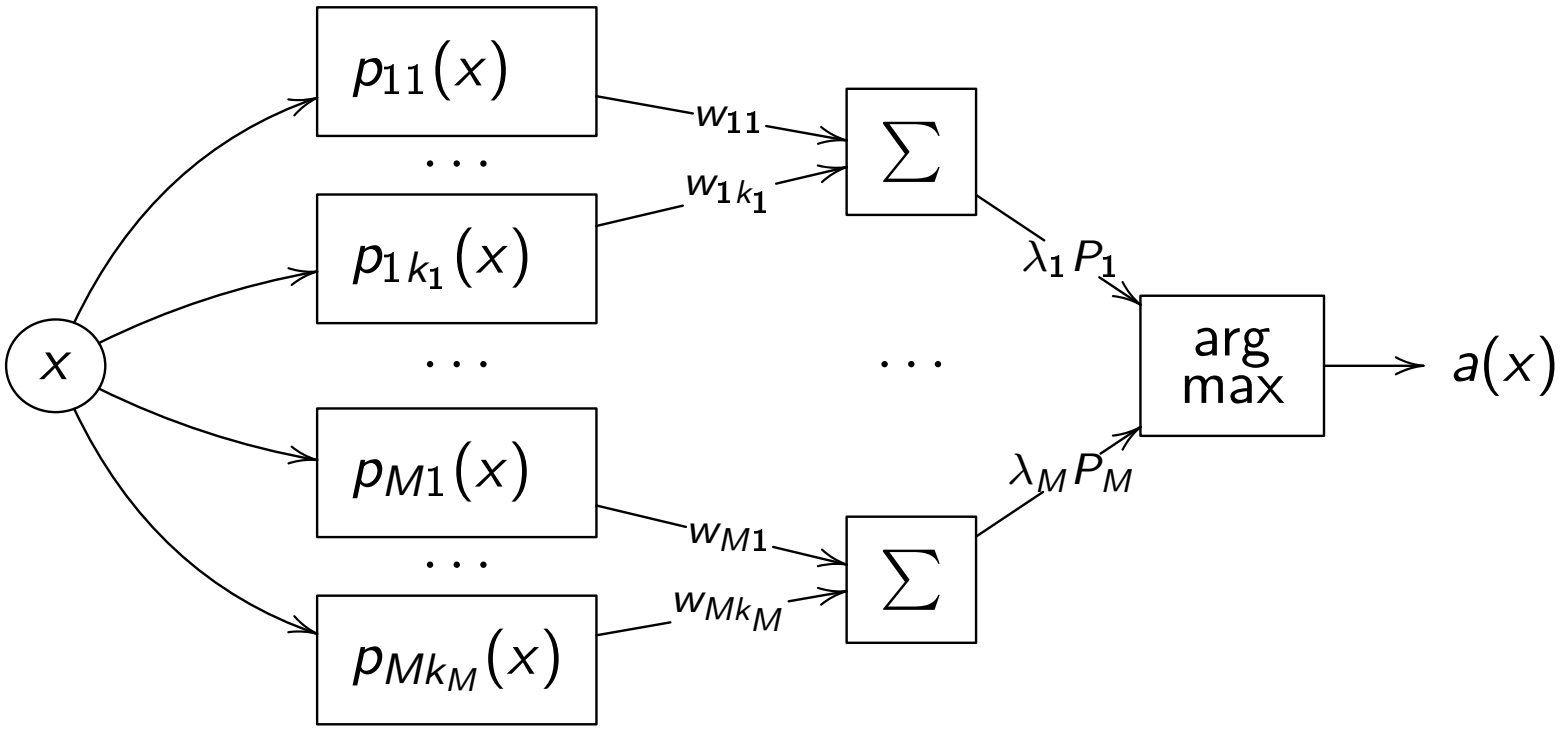
$$\rho^2_{yj}(x, \mu_{yj}) = \sum_{d=1}^{n} \frac{1}{\sigma^2_{yjd}} \big(f_d(x) - \mu_{yjd}\big)^2.$$

Интерпретация:
$p_{yj}(x)$ — близость объекта x к j-й компоненте класса y;
$\Gamma_y(x)$ — близость объекта x к классу y.

Сеть радиальных базисных функций (RBF)

Трёхслойная сеть RBF (Radial Basis Functions):

$$a(x) = \arg\max_{y \in Y} \lambda_y P_y \sum_{j=1}^{k_y} w_{yj} p_{yj}(x)$$



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
- Два подхода к обучению классификации:
  — *дискриминативный*: модель вероятности классов $P(y|x,w)$ (LSM, LR, GLM, SVM, RBF,...)
  — *генеративный*: модель плотности классов $p(x|y,w)$ (OBC: NB, PW, FLD, RBF,...)
- *Байесовское обучение:* $p(w|X^\ell) = \dfrac{p(X^\ell|w)p(w)}{\int p(X^\ell|w)p(w)\,dw}$