Статистические (байесовские) методы классификации

К. В. Воронцов
vokov@forecsys.ru

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

февраль 2013

Содержание

Постановка задачи

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

Дано:

$X^\ell = (x_i, y_i)_{i=1}^\ell$ — простая выборка;

Найти:

классификатор $a\colon X \to Y$ с минимальной вероятностью ошибки.

Временное допущение: пусть известна совместная плотность

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

$\mathrm{P}(y) \equiv P_y$ — *априорная вероятность* класса y;

$p(x|y) \equiv p_y(x)$ — *функция правдоподобия* класса y;

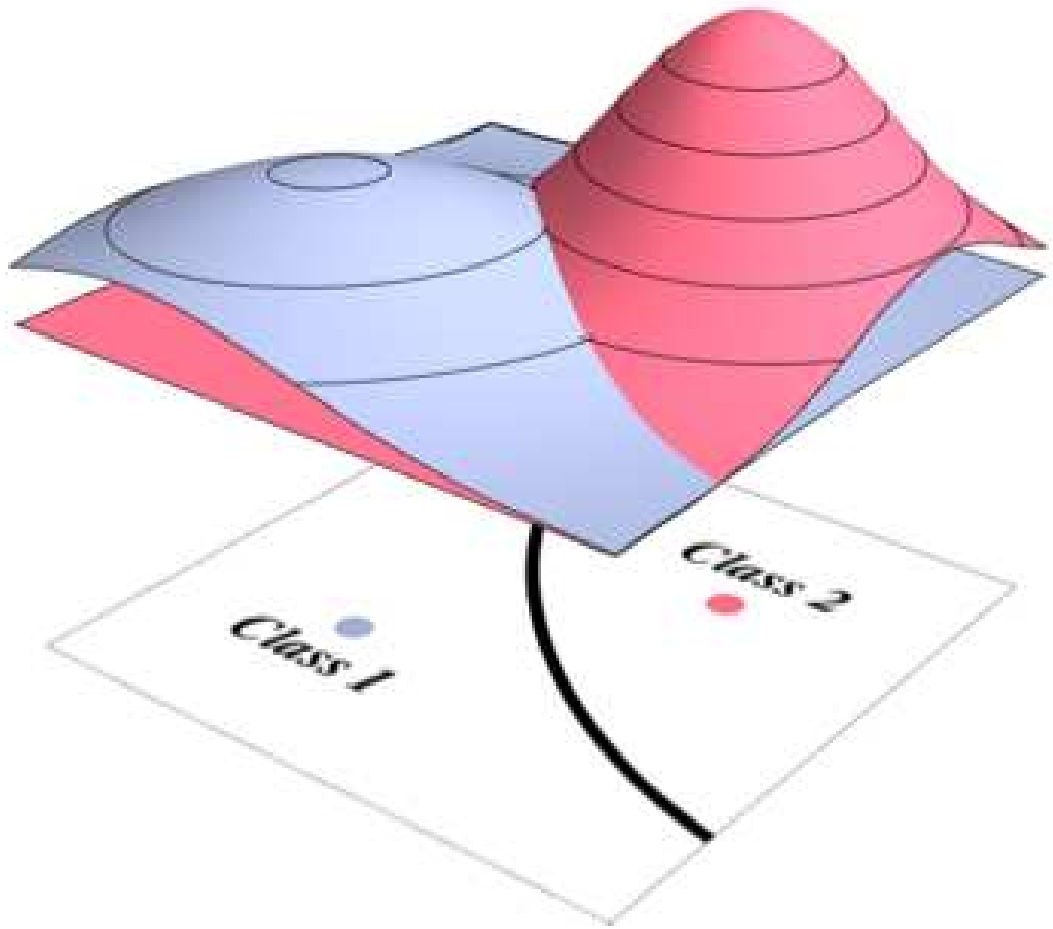
$\mathrm{P}(y|x)$ — *апостериорная вероятность* класса y;

Принцип максимума апостериорной вероятности:

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

Классификация по максимуму функции правдоподобия

Частный случай: $a(x) = \arg\max\limits_{y \in Y} p_y(x)$ при $P_y = \text{const}$.



Функционал среднего риска

$a\colon X \to Y$ разбивает X на непересекающиеся области:

$$A_y = \{x \in X \mid a(x) = y\}, \quad y \in Y.$$

Ошибка: объект x класса y попадает в A_s, $s \neq y$.

Вероятность ошибки: $\mathsf{P}(A_s, y) = \int_{A_s} p(x,y)dx$.

Потеря от ошибки: задана $\lambda_{ys} \geqslant 0$, для всех $(y,s) \in Y \times Y$.

Средний риск — мат.ожидание потери для классификатора a:

$$R(a) = \sum_{y \in Y} \sum_{s \in Y} \lambda_{ys} \mathsf{P}(A_s, y),$$

Две теоремы об оптимальности байесовского классификатора

Теорема

Если известны $P_y = \mathrm{P}(y)$ и $p_y(x) = p(x|y)$, то минимум среднего риска $R(a)$ достигается при

$$a(x) = \arg\min_{s \in Y} \sum_{y \in Y} \lambda_{ys} P_y p_y(x).$$

Теорема

Если к тому же $\lambda_{yy} = 0$ и $\lambda_{ys} \equiv \lambda_y$ для всех $y, s \in Y$, то минимум среднего риска $R(a)$ достигается при

$$a(x) = \arg\max_{y \in Y} \lambda_y P_y p_y(x).$$

При чём тут Байес?

Апостериорная вероятность по формуле Байеса:

$$\mathrm{P}(y|x) = \frac{p(x,y)}{p(x)} = \frac{P_y p_y(x)}{\sum\limits_{s \in Y} P_s p_s(x)}.$$

Если $\lambda_y = 1$, то получаем всё тот же
принцип максимума апостериорной вероятности:

$$a(x) = \arg\max_{y \in Y} \lambda_y P_y p_y(x) = \arg\max_{y \in Y} \mathrm{P}(y|x).$$

Ожидаемая потеря на объекте x:

$$R(x) = \sum_{y \in Y} \lambda_y \mathrm{P}(y|x).$$

Итак, есть две подзадачи, причём вторую мы уже решили!

1. **Дано:**
   $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка.
   Найти:
   эмпирические оценки $\hat{P}_y$ и $\hat{p}_y(x)$, $y \in Y$
   (восстановить плотность распределения по выборке).
2. **Дано:**
   априорные вероятности P_y,
   функции правдоподобия $p_y(x)$, $y \in Y$.
   Найти:
   классификатор $a\colon X \times Y$, минимизирующий $R(a)$.

Ехидное замечание: Когда вместо P_y и $p_y(x)$ подставляются их эмпирические оценки, байесовский классификатор перестаёт быть оптимальным.

Задачи эмпирического оценивания

- **Оценивание априорных вероятностей** частотами

$$\hat{P}_y = \frac{\ell_y}{\ell}, \quad \ell_y = |X_y|, \quad X_y = \{x_i \in X : y_i = y\}, \quad y \in Y.$$

- **Оценивание функций правдоподобия:**
  Дано:
  $X^m = \{x_1, \ldots, x_m\}$ — простая выборка (X_y без ответов y_i).
  Найти:
  эмпирическую оценку плотности $\hat{p}(x)$,
  аппроксимирующую истинную плотность $p(x)$ на всём X:

$$\hat{p}(x) \to p(x) \text{ при } m \to \infty.$$

Анонс: три подхода к оцениванию плотностей

1. **Параметрическое оценивание плотности:**

$$\hat{p}(x) = \varphi(x, \theta).$$

2. **Восстановление смеси распределений:**

$$\hat{p}(x) = \sum_{j=1}^{k} w_j \varphi(x, \theta_j), \quad k \ll m.$$

3. **Непараметрическое оценивание плотности:**

$$\hat{p}(x) = \sum_{i=1}^{m} \frac{1}{mV(h)} K\left(\frac{\rho(x, x_i)}{h}\right).$$

Наивный байесовский классификатор

Допущение (наивное):

Признаки $f_j : X \to D_j$ — независимые случайные величины с плотностями распределения, $p_{y,j}(\xi)$, $y \in Y$, $j = 1, \ldots, n$.

Тогда функции правдоподобия классов представимы в виде произведения одномерных плотностей по признакам:

$$p_y(x) = p_{y,1}(\xi_1) \cdots p_{y,n}(\xi_n), \quad x = (\xi_1, \ldots, \xi_n), \quad y \in Y.$$

Прологарифмируем (для удобства). Получим классификатор

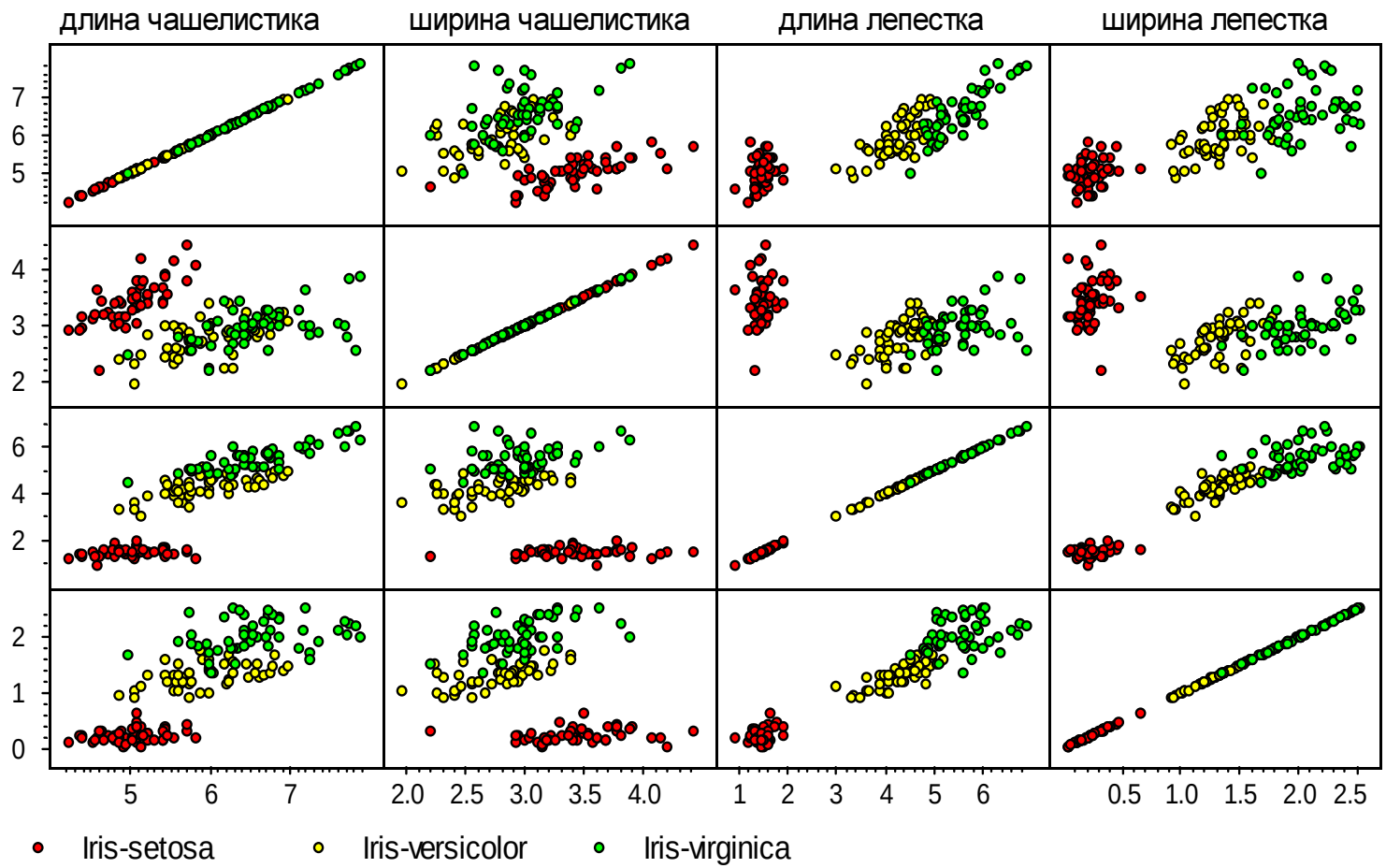
$$a(x) = \arg\max_{y \in Y} \left(\ln \lambda_y \hat{P}_y + \sum_{j=1}^{n} \ln \hat{p}_{yj}(\xi_j) \right).$$

Восстановление n одномерных плотностей
— намного более простая задача, чем одной n-мерной.

Ирисы Фишера

$n = 4$ признака, $|Y| = 3$ класса, длина выборки $\ell = 150$.

Можно ли здесь применить наивный байесовский классификатор?



Начнём с определения плотности вероятности

Дискретный случай: $|X| \ll m$. Гистограмма значений x_i:

$$\hat{p}(x) = \frac{1}{m} \sum_{i=1}^{m} [x_i = x].$$

Одномерный непрерывный случай: $X = \mathbb{R}$. По определению плотности, если $P[a, b]$ — вероятностная мера отрезка $[a, b]$:

$$p(x) = \lim_{h \to 0} \frac{1}{2h} P[x - h, x + h],$$

Эмпирическая оценка плотности по окну ширины h:

$$\hat{p}_h(x) = \frac{1}{2mh} \sum_{i=1}^{m} \big[|x - x_i| < h\big].$$

Локальная непараметрическая оценка Парзена-Розенблатта

Эмпирическая оценка плотности по окну ширины h:

$$\hat{p}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} \frac{1}{2} \left[\frac{|x - x_i|}{h} < 1 \right].$$

Обобщение: *оценка Парзена-Розенблатта* по окну ширины h:

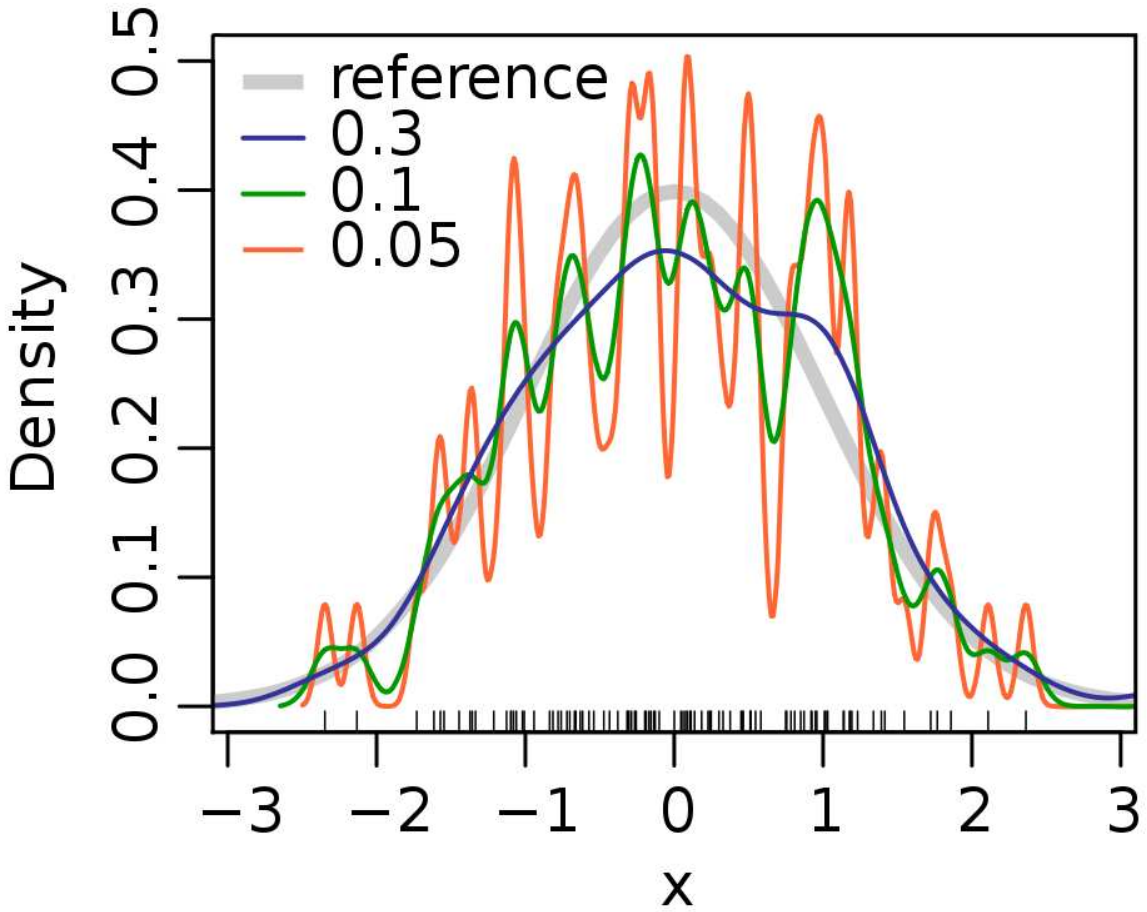
$$\hat{p}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h} \right),$$

где $K(r)$ — *ядро*, удовлетворяющее требованиям:
— чётная функция;
— нормированная функция: $\int K(r)\, dr = 1$;
— (как правило) невозрастающая, неотрицательная функция.

В частности, при $K(r) = \frac{1}{2}\big[|r| < 1\big]$ имеем эмпирическую оценку.

Пример. Ядерные оценки плотности при разных h

Оценка $\hat{p}_h(x)$ существенно зависит от ширины окна h:



Обоснование оценки Парзена-Розенблатта

Теорема (одномерный случай, $X = \mathbb{R}$)

Пусть выполнены следующие условия:

1) X^m — простая выборка из распределения $p(x)$;
2) ядро $K(z)$ непрерывно и ограничено: $\int_X K^2(z)\,dz < \infty$;
3) последовательность h_m: $\lim\limits_{m\to\infty} h_m = 0$ и $\lim\limits_{m\to\infty} mh_m = \infty$.

Тогда:

1) $\hat{p}_{h_m}(x) \to p(x)$ при $m \to \infty$ для почти всех $x \in X$;
2) скорость сходимости имеет порядок $O(m^{-2/5})$.

А как быть в многомерном случае, когда $X = \mathbb{R}^n$?

Два варианта обобщения на многомерный случай

1. Если объекты описываются n числовыми признаками $f_j : X \to \mathbb{R},\ j = 1, \ldots, n$.

$$\hat{p}_h(x) = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{n} \frac{1}{h_j} K\left(\frac{f_j(x) - f_j(x_i)}{h_j}\right).$$

2. Если на X задана функция расстояния $\rho(x, x')$:

$$\hat{p}_h(x) = \frac{1}{mV(h)} \sum_{i=1}^{m} K\left(\frac{\rho(x, x_i)}{h}\right),$$

где $V(h) = \int_X K\left(\frac{\rho(x,x_i)}{h}\right) dx$ — нормирующий множитель.

Замечание: $V(h)$ не должен зависеть от x_i (однородность $\langle X, \rho \rangle$).

Упражнение: Приведите примеры таких K и ρ, чтобы варианты 1 и 2 оказались эквивалентными.

Метод парзеновского окна

Парзеновская оценка плотности для каждого класса $y \in Y$:

$$\hat{p}_{y,h}(x) = \frac{1}{\ell_y V(h)} \sum_{i:\, y_i = y} K\left(\frac{\rho(x, x_i)}{h}\right),$$

Метод парзеновского окна (Parzen window):

$$a(x; X^\ell, h) = \arg\max_{y \in Y} \lambda_y \frac{P_y}{\ell_y} \sum_{i:\, y_i = y} K\left(\frac{\rho(x, x_i)}{h}\right).$$

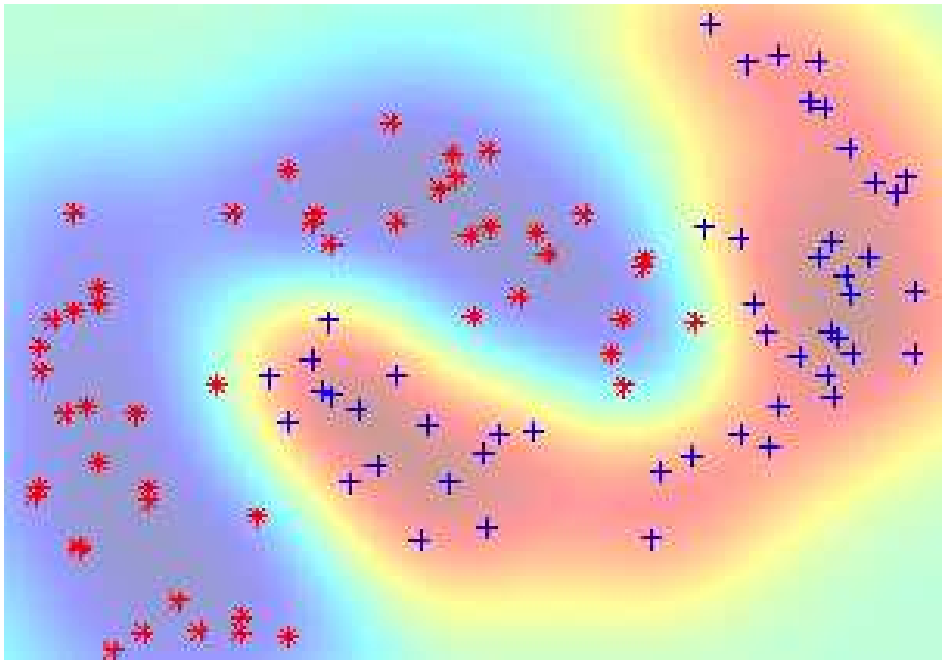
Остаются вопросы:
1) на что влияет ядро $K(r)$ и как его выбрать?
2) на что влияет ширина окна h и как её выбрать?
3) откуда взять функцию расстояния $\rho(x, x')$?

Пример. Визуализация парзеновского классификатора

Метод парзеновского окна (Parzen window):

$$a(x; X^\ell, h) = \arg\max_{y \in Y} \Gamma_y(x), \quad \Gamma_y(x) = \lambda_y \frac{P_y}{\ell_y} \sum_{i:\, y_i = y} K\left(\frac{\rho(x, x_i)}{h}\right)$$

Цветом передаётся значение разности $\Gamma_+(x) - \Gamma_*(x)$:



Выбор метрики (функция расстояния)

Один из возможных вариантов
— взвешенная метрика Минковского:

$$\rho(x, x') = \left(\sum_{j=1}^{n} w_j \left| f_j(x) - f_j(x') \right|^p \right)^{\frac{1}{p}},$$

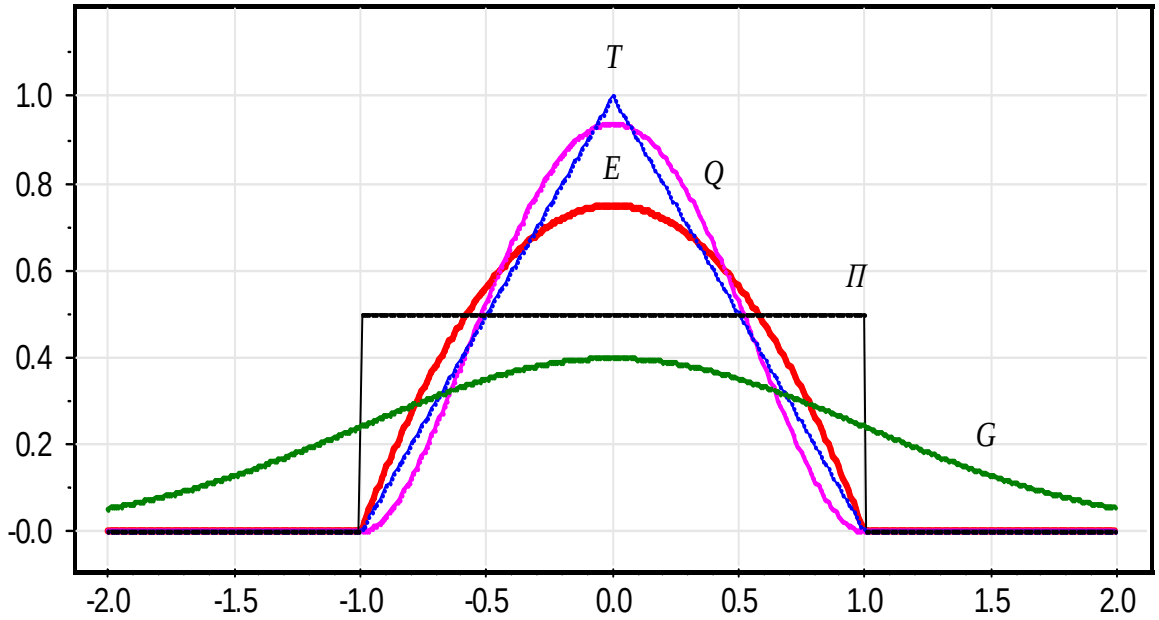
где w_j — неотрицательные веса признаков, $p > 0$.

В частности, если $w_j \equiv 1$ и $p = 2$, то имеем евклидову метрику.

Роль весов w_j:

1) нормировка признаков;
2) степень важности признаков;
3) отбор признаков (какие $w_j = 0$?);

Часто используемые ядра



$E(r) = \frac{3}{4}(1-r^2)\big[|r| \leqslant 1\big]$ — оптимальное (Епанечникова);
$Q(r) = \frac{15}{16}(1-r^2)^2\big[|r| \leqslant 1\big]$ — квартическое;
$T(r) = \big(1-|r|\big)\big[|r| \leqslant 1\big]$ — треугольное;
$G(r) = (2\pi)^{-1/2}\exp(-\frac{1}{2}r^2)$ — гауссовское;
$\Pi(r) = \frac{1}{2}\big[|r| \leqslant 1\big]$ — прямоугольное.

Выбор ядра почти не влияет на качество восстановления

Функционал качества восстановления плотности:

$$J(K) = \int_{-\infty}^{+\infty} \mathsf{E}\big(\hat{p}_h(x) - p(x)\big)^2\, dx.$$

ядро $K(r)$	степень гладкости	$J(K^*)/J(K)$
Епанечникова $K^*(r)$	$\hat{p}'_h$ разрывна	1.000
Квартическое	$\hat{p}''_h$ разрывна	0.995
Треугольное	$\hat{p}'_h$ разрывна	0.989
Гауссовское	∞ дифференцируема	0.961
Прямоугольное	$\hat{p}_h$ разрывна	0.943

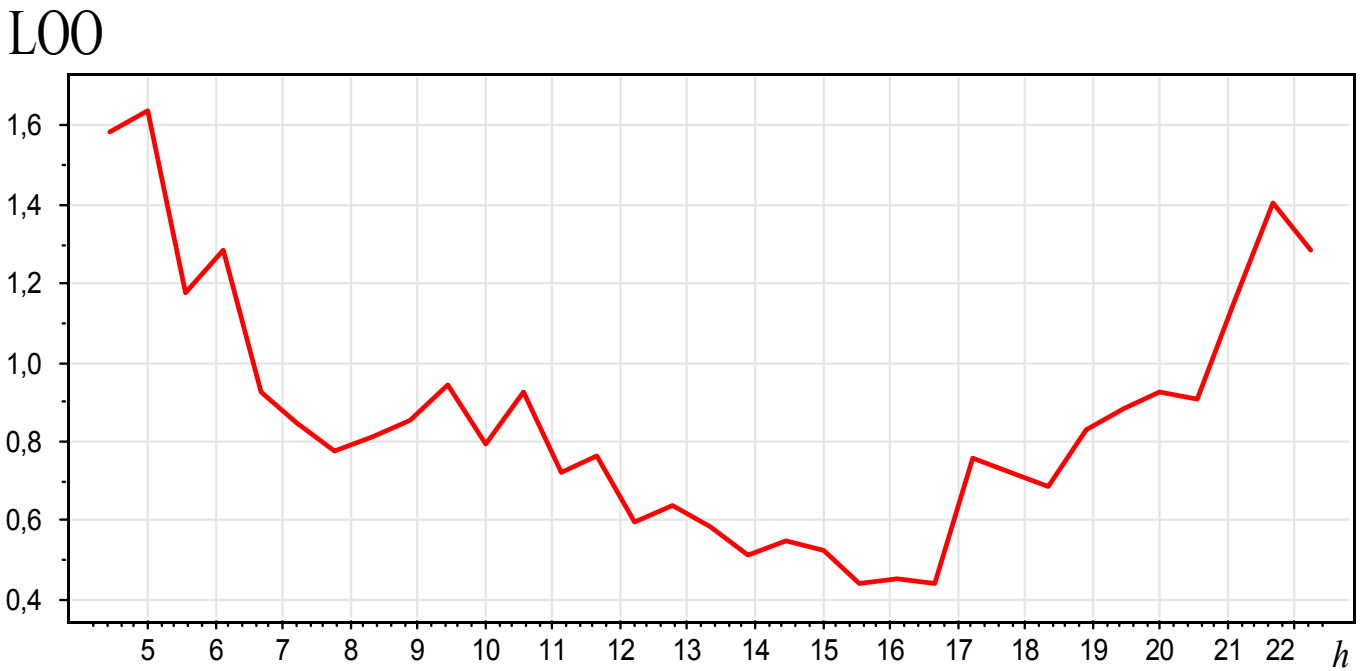
Замечание: в таблице представлены асимптотические значения отношения $J(K^*)/J(K)$ при $m \to \infty$, причём это отношение не зависит от $p(x)$.

Выбор ширины окна

Скользящий контроль *Leave One Out*:

$$\mathrm{LOO}(h, X^\ell) = \sum_{i=1}^{\ell} \Big[a\big(x_i; X^\ell \setminus x_i, h\big) \neq y_i \Big] \to \min_h,$$

Типичный вид зависимости LOO от h:



Окна переменной ширины

Проблема:

при наличии локальных сгущений любая h не оптимальна.

Идея:

задавать не ширину окна h, а число соседей k.

$$h(x) = \rho\big(x, x^{(k+1)}\big),$$

где $x^{(i)}$ — i-й сосед объекта x при ранжировании выборки X^ℓ:

$$\rho(x, x^{(1)}) \leqslant \cdots \leqslant \rho(x, x^{(\ell)})$$

Замечание 1:

нормировка $V(k)$ не должна зависеть от y, поэтому выборка ранжируется целиком, а не по классам X_y.

Замечание 2:

Оптимизация k по LOO аналогична оптимизации h.

Резюме в конце лекции

- $a(x) = \arg\max_{y \in Y} \lambda_y P_y p_y(x)$ — эту формулу надо помнить.
- Наивный байесовский классификатор основан на «драконовском» предположении о независимости признаков. Как ни странно, иногда это работает.
- Три основных подхода к восстановлению функций правдоподобия $p_y(x)$ по выборке: параметрический, непараметрический и смесь распределений.
- Непараметрический подход наиболее прост и приводит к методу парзеновского окна.
- Проблемы непараметрического подхода:
  — выбор ширины окна h или числа соседей k;
  — выбор сглаживающего ядра K;
  — выбор метрики.

Принцип максимума правдоподобия

Пусть известна параметрическая модель плотности

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

где θ — параметр, φ — фиксированная функция.

Задача — найти оптимальное θ по простой выборке X^m.

Принцип максимума (взвешенного) правдоподобия:

$$L(\theta; X^m, G^m) = \sum_{i=1}^{m} g_i \ln \varphi(x_i;\theta) \to \max_{\theta},$$

где $G^m = (g_1, \ldots, g_m)$ — вектор весов объектов.

Необходимое условие оптимума:

$$\frac{\partial}{\partial\theta} L(\theta; X^m, G^m) = \sum_{i=1}^{m} g_i \frac{\partial}{\partial\theta} \ln \varphi(x_i;\theta) = 0,$$

где функция $\varphi(x;\theta)$ достаточно гладкая по параметру θ.

Многомерное нормальное распределение

Пусть $X = \mathbb{R}^n$ — объекты описываются n числовыми признаками.

Гипотеза: классы имеют n-мерные гауссовские плотности:

$$p_y(x) = \mathcal{N}(x; \mu_y, \Sigma_y) = \frac{e^{-\frac{1}{2}(x-\mu_y)^\intercal \Sigma_y^{-1}(x-\mu_y)}}{\sqrt{(2\pi)^n \det \Sigma_y}}, \quad y \in Y,$$

где $\mu_y \in \mathbb{R}^n$ — вектор матожидания (центр) класса $y \in Y$,
$\Sigma_y \in \mathbb{R}^{n \times n}$ — ковариационная матрица класса $y \in Y$
(симметричная, невырожденная, положительно определённая).

Теорема

1. Разделяющая поверхность $\left\{x \in X \;\middle|\; \lambda_y P_y p_y(x) = \lambda_s P_s p_s(x)\right\}$ квадратична для всех $y, s \in Y$, $y \neq s$.
2. Если $\Sigma_y = \Sigma_s$, то она вырождается в линейную.

Квадратичный дискриминант

Теорема

Оценки максимума взвешенного правдоподобия, $y \in Y$:

$$\hat{\mu}_y = \frac{1}{G_y} \sum_{i:\, y_i = y} g_i x_i;$$

$$\hat{\Sigma}_y = \frac{1}{G_y} \sum_{i:\, y_i = y} g_i (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^\intercal;$$

где $G_y = \sum_{i:\, y_i = y} g_i.$

Квадратичный дискриминант — подстановочный алгоритм:

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

Квадратичный дискриминант

Недостатки квадратичного дискриминанта:

- Если $\ell_y < n$, то матрица $\hat{\Sigma}_y$ вырождена.
- Чем меньше ℓ_y, тем менее устойчива оценка $\hat{\Sigma}_y$.
- Оценки $\hat{\mu}_y$, $\hat{\Sigma}_y$ неустойчивы к шуму.
- Если классы не нормальны, всё совсем плохо...

Меры по улучшению алгоритма:

- Линейный дискриминант (вместо квадратичного)
- Регуляризация ковариационной матрицы
- Цензурирование выборки (отсев шума)
- Смеси нормальных распределений

Линейный дискриминант Фишера

Допущение:

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

Линейный дискриминант — подстановочный алгоритм:

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

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

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

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

Недостаток: всё равно приходится обращать матрицу $\hat{\Sigma}$.

Проблема мулитиколлинеарности

Мулитиколлинеарность
— это когда матрица $\hat{\Sigma}$ близка к вырожденной.

Проявления мулитиколлинеарности:
1) некоторые собственные значения $\hat{\Sigma}$ близки к нулю;
2) обратная $\hat{\Sigma}^{-1}$ неустойчива;
3) нормаль разделяющей гиперплоскости $\alpha_y = \hat{\Sigma}^{-1}\hat{\mu}_y$ неустойчива;
4) переобучение: на X^ℓ всё хорошо, на X^k всё плохо.

Пути повышения качества классификации

- Регуляризация ковариационной матрицы
- Обнуление элементов ковариационной матрицы
- Диагонализация ковариационной матрицы
- Понижение размерности
- Редукция размерности по А.М.Шурыгину
- Цензурирование выборки (отсев шума)
- Усложнение модели (смесь нормальных распределений)

Регуляризация ковариационной матрицы

Идея:

преобразовать матрицу $\hat{\Sigma}$ так, чтобы
все собственные векторы v остались,
а все собственные значения λ увеличились на τ:

$$(\hat{\Sigma} + \tau I_n)v = \lambda v + \tau v = (\lambda + \tau)v.$$

Рецепт:

1) обращаем $\hat{\Sigma} + \tau I_n$ вместо $\hat{\Sigma}$;

2) параметр регуляризации τ подбираем по скользящему контролю.

Обнуление элементов ковариационной матрицы

$\hat{\Sigma} = \|\sigma_{ij}\|_{n\times n}$

Идея: обнулить статистически незначимые ковариации σ_{ij}.

Воплощение:
Для всех $i, j = 1, \ldots, n, \;\; i < j$

1) вычисляется коэффициент корреляции $r_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$;

2) статистика $T_{ij} = \dfrac{r_{ij}\sqrt{n-2}}{\sqrt{1-r_{ij}^2}}$ имеет
t-распределение Стьюдента с $n-2$ степенями свободы;

3) если $|T_{ij}| \leqslant t_{1-\frac{\alpha}{2}}$ — квантиль распределения Стьюдента при заданном уровне значимости α,
то полагается $\sigma_{ij} := \sigma_{ji} := 0$.

Диагонализация ковариационной матрицы

Идея: пусть признаки некоррелированы: $\sigma_{ij} = 0, \quad i \neq j$.

Замечание: для нормального распределения
некоррелированность $\Longleftrightarrow$ независимость

Получаем **наивный байесовский классификатор:**

$$\hat{p}_{yj}(\xi) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_{yj}} \exp\left(-\frac{(\xi - \hat{\mu}_{yj})^2}{2\hat{\sigma}_{yj}^2}\right), \quad y \in Y, \quad j = 1, \ldots, n;$$

$$a(x) = \arg\max_{y \in Y}\left(\ln \lambda_y \hat{P}_y + \sum_{j=1}^{n} \ln \hat{p}_{yj}(\xi_j)\right), \quad x \equiv (\xi_1, \ldots, \xi_n);$$

где $\hat{\mu}_{yj}$ и $\hat{\sigma}_{yj}$ — оценки среднего и дисперсии j-го признака, вычисленные по X_y — подвыборке класса y.

Понижение размерности

Идея 1:
отбор признаков (features selection)

Идея 2:
преобразование n признаков в $m < n$ признаков (PCA)

Эти подходы будут разбираться в следующих лекциях.

Редукция размерности по А. М. Шурыгину

Идея:

сведение n-мерной задачи к серии двумерных задач путём подключения признаков по одному.

Набросок алгоритма:

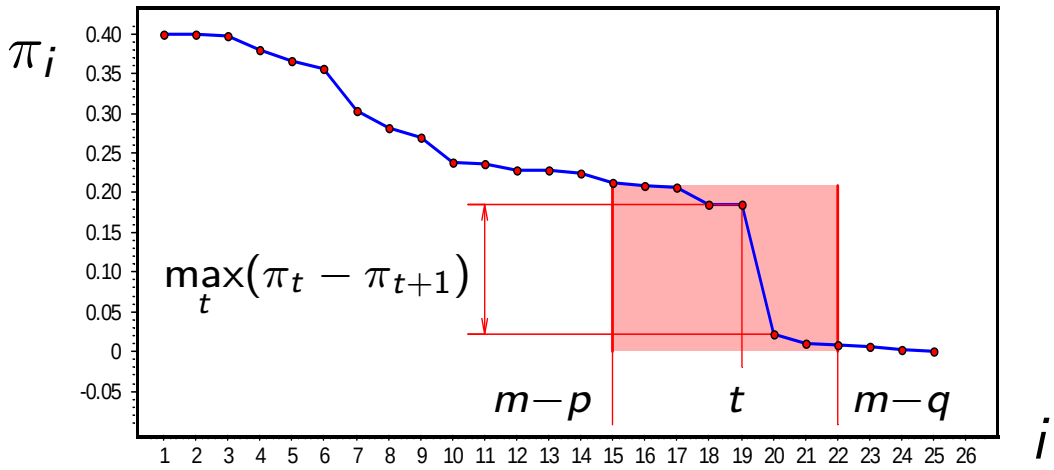
1) найти два признака, в подпространстве которых классы наилучшим образом разделимы;
2) новый признак: $\psi(x) = x^{\mathsf{T}}\alpha_y$ — проекция на нормаль к разделяющей прямой в пространстве двух признаков;
3) выбрать из оставшихся признаков тот, который в паре с $\psi(x)$ даёт наилучшую разделимость;
4) если разделимость не улучшилась, прекратить;
5) иначе GOTO 2);

Цензурирование выборки (отсев шума)

Идея: задача решается дважды; после первого раза объекты с наибольшими ошибками исключаются из обучения.

Алгоритм (для задачи восстановления плотности)

1) оценить параметр $\hat{\theta}$ по всей выборке X^m;
2) вычислить правдоподобия $\pi_i = \varphi(x_i; \hat{\theta})$ для всех $x_i \in X^m$;
3) отсортировать выборку по убыванию: $\pi_1 \geqslant \ldots \geqslant \pi_m$;
4) удалить из X^m объекты, попавшие в конец ряда;
5) оценить параметр $\hat{\theta}$ по укороченной выборке X^m;



Резюме в конце лекции

- Параметрический подход = модель плотности распределения + принцип максимума правдоподобия.
- Модель гауссовских плотностей приводит к квадратичному или линейному дискриминанту.
- Их основная проблема — неустойчивость обращения ковариационной матрицы. Способы решения:
  — регуляризация;
  — диагонализация;
  — обнуление незначимых ковариаций;
  — снижение размерности путём отбора признаков;
  — жадное добавление признаков (метод Шурыгина);
  — снижение размерности путём преобразования признаков.

Модель смеси распределений

Модель плотности:

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

$p_j(x) = \varphi(x; \theta_j)$ — *функция правдоподобия* j-й компоненты смеси;
w_j — её *априорная вероятность*; k — число компонент смеси.

Задача 1: имея простую выборку $X^m \sim p(x)$,
зная число k и функцию φ, оценить вектор параметров
$\Theta = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$.

Задача 2: оценить ещё и k.

Общая схема EM-алгоритма

Проблема:
попытка применить принцип максимума правдоподобия «в лоб» приводит к очень сложной многоэкстремальной задаче оптимизации

Идея: вводятся *скрытые переменные* G.

Итерационный алгоритм Expectation–Maximization:

1: начальное приближение вектора параметров Θ;
2: **повторять**
3: $\quad G := \text{E-шаг}\,(\Theta)$;
4: $\quad \Theta := \text{M-шаг}\,(\Theta, G)$;
5: **пока** Θ и G не стабилизируются.

Задача E-шага

По формуле условной вероятности

$$p(x, \theta_j) = p(x)\,\mathrm{P}(\theta_j|x) = w_j p_j(x).$$

Скрытые переменные $G = (g_{ij})_{m\times k} = (g_1, \ldots, g_j)$:

$$g_{ij} \equiv \mathrm{P}(\theta_j|x_i), \quad i = 1, \ldots, m,\ j = 1, \ldots, k.$$

Зная параметры компонент w_j, θ_j, по формуле Байеса легко вычислить g_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k$:

$$g_{ij} = \frac{w_j p_j(x_i)}{p(x_i)} = \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)}.$$

Очевидно, выполнено условие нормировки: $\sum_{j=1}^{k} g_{ij} = 1$.

Задача M-шага

Задача: максимизировать логарифм правдоподобия

$$Q(\Theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i) \to \max_{\Theta}.$$

при ограничениях $\sum_{j=1}^{k} w_j = 1;\ \ w_j \geqslant 0.$

Если скрытые переменные известны, то задача максимизации $Q(\Theta)$ распадается на k независимых подзадач:

$$\theta_j := \arg\max_{\theta} \sum_{i=1}^{m} g_{ij} \ln \varphi(x_i; \theta), \quad j = 1, \ldots, k.$$

а оптимальные веса компонент вычисляются аналитически:

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

Вывод формул M-шага (основные шаги)

Лагранжиан оптимизационной задачи «$Q(\Theta) \to \max$»:

$$L(\Theta; X^m) = \sum_{i=1}^{m} \ln \Bigg(\underbrace{\sum_{j=1}^{k} w_j p_j(x_i)}_{p(x_i)} \Bigg) - \lambda \Bigg(\sum_{j=1}^{k} w_j - 1 \Bigg).$$

Приравниваем нулю производные:

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

$$\frac{\partial L}{\partial \theta_j} = \sum_{i=1}^{m} \underbrace{\frac{w_j p_j(x_i)}{p(x_i)}}_{g_{ij}} \frac{\partial}{\partial \theta_j} \ln p_j(x_i) = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{m} g_{ij} \ln p_j(x_i) = 0.$$

EM-алгоритм

Вход:

выборка $X^m = \{x_1, \ldots, x_m\}$;

k — число компонент смеси;

$\Theta = (w_j, \theta_j)_{j=1}^k$ — начальное приближение параметров;

δ — параметр критерия останова;

Выход:

$\Theta = (w_j, \theta_j)_{j=1}^k$ — оптимизированный вектор параметров для смеси $p(x) = \sum\limits_{j=1}^{k} w_j \varphi(x, \theta_j)$, $\sum\limits_{j=1}^{k} w_j = 1$.

Базовый вариант EM-алгоритма

1: **ПРОЦЕДУРА** EM (X^m, k, Θ, δ);
2: **повторять**
3: E-шаг (expectation):
для всех $i = 1, \ldots, m,\ \ j = 1, \ldots, k$

$$g_{ij}^0 := g_{ij}; \qquad g_{ij} := \frac{w_j \varphi(x_i; \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i; \theta_s)};$$

4: M-шаг (maximization):
для всех $j = 1, \ldots, k$

$$\theta_j := \arg\max_{\theta} \sum_{i=1}^m g_{ij} \ln \varphi(x_i; \theta); \qquad w_j := \frac{1}{m} \sum_{i=1}^m g_{ij};$$

5: **пока** $\max\limits_{i,j} |g_{ij} - g_{ij}^0| > \delta$;
6: **вернуть** $(w_j, \theta_j)_{j=1}^k$;

Проблемы базового варианта EM-алгоритма

- Как выбирать начальное приближение?
- Какой выбрать критерий останова?
- Как определять число компонент?
- Как ускорить сходимость?

Решение сразу многих проблем:
EM-алгоритм с последовательным добавлением компонент

EM-алгоритм с последовательным добавлением компонент

Вход:

выборка $X^m = \{x_1, \ldots, x_m\}$;

R — допустимый разброс правдоподобия объектов;

m_0 — минимальная длина выборки, по которой можно восстанавливать плотность;

δ — параметр критерия останова;

Выход:

k — число компонент смеси;

$\Theta = (w_j, \theta_j)_{j=1}^k$ — веса и параметры компонент;

EM-алгоритм с последовательным добавлением компонент

1: начальное приближение — одна компонента:
$\theta_1 := \arg\max_\theta \sum_{i=1}^{m} \ln \varphi(x_i;\theta)$; $w_1 := 1$; $k := 1$;
2: **для всех** $k := 2, 3, \ldots$
3: выделить объекты с низким правдоподобием:
$U := \left\{ x_i \in X^m \;\middle|\; p(x_i) < \frac{1}{R} \max_j p(x_j) \right\}$;
4: **если** $|U| < m_0$ **то**
5: **выход** из цикла по k;
6: начальное приближение для k-й компоненты:
$\theta_k := \arg\max_\theta \sum_{x_i \in U} \ln \varphi(x_i;\theta)$; $w_k := \frac{1}{m}|U|$;
$w_j := w_j(1 - w_k)$, $j = 1, \ldots, k-1$;
7: выполнить EM (X^m, k, Θ, δ);

GEM — обобщённый EM-алгоритм

Идея:
Не обязательно добиваться высокой точности на M-шаге.
Достаточно лишь сместиться в направлении максимума,
сделав одну или несколько итераций, и затем выполнить E-шаг.

Преимущество:
уменьшение времени работы при сопоставимом качестве решения.

SEM — стохастический EM-алгоритм

Идея: на M-шаге вместо максимизации

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

максимизируется обычное, невзвешенное, правдоподобие

$$\theta_j := \arg\max_{\theta} \sum_{x_i \in X_j} \ln \varphi(x_i; \theta),$$

выборки X_j строятся путём стохастического моделирования:
для каждого $i = 1, \ldots, m$ генерируется $j(i) \in \{1, \ldots, k\}$:
$\mathsf{P}\{j(i) = s\} = g_{is}$, и объект x_i помещается в $X_{j(i)}$.

Преимущества:
ускорение сходимости, предотвращение зацикливаний.

HEM — иерархический EM-алгоритм

Идея:
«Плохо описанные» компоненты расщепляются на две или более *дочерних* компонент.

Преимущество:
автоматически выявляется иерархическая структура каждого класса, которую затем можно интерпретировать содержательно.

Гауссовская смесь с диагональными матрицами ковариации

Допущения:

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

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

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

Эмпирические оценки средних и дисперсий

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

Решение задачи M-шага:

для всех классов $y \in Y$ и всех компонент $j = 1, \ldots, k_y$,

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

для всех размерностей (признаков) $d = 1, \ldots, n$

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

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

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

Алгоритм классификации

Подставим гауссовскую смесь в байесовский классификатор:

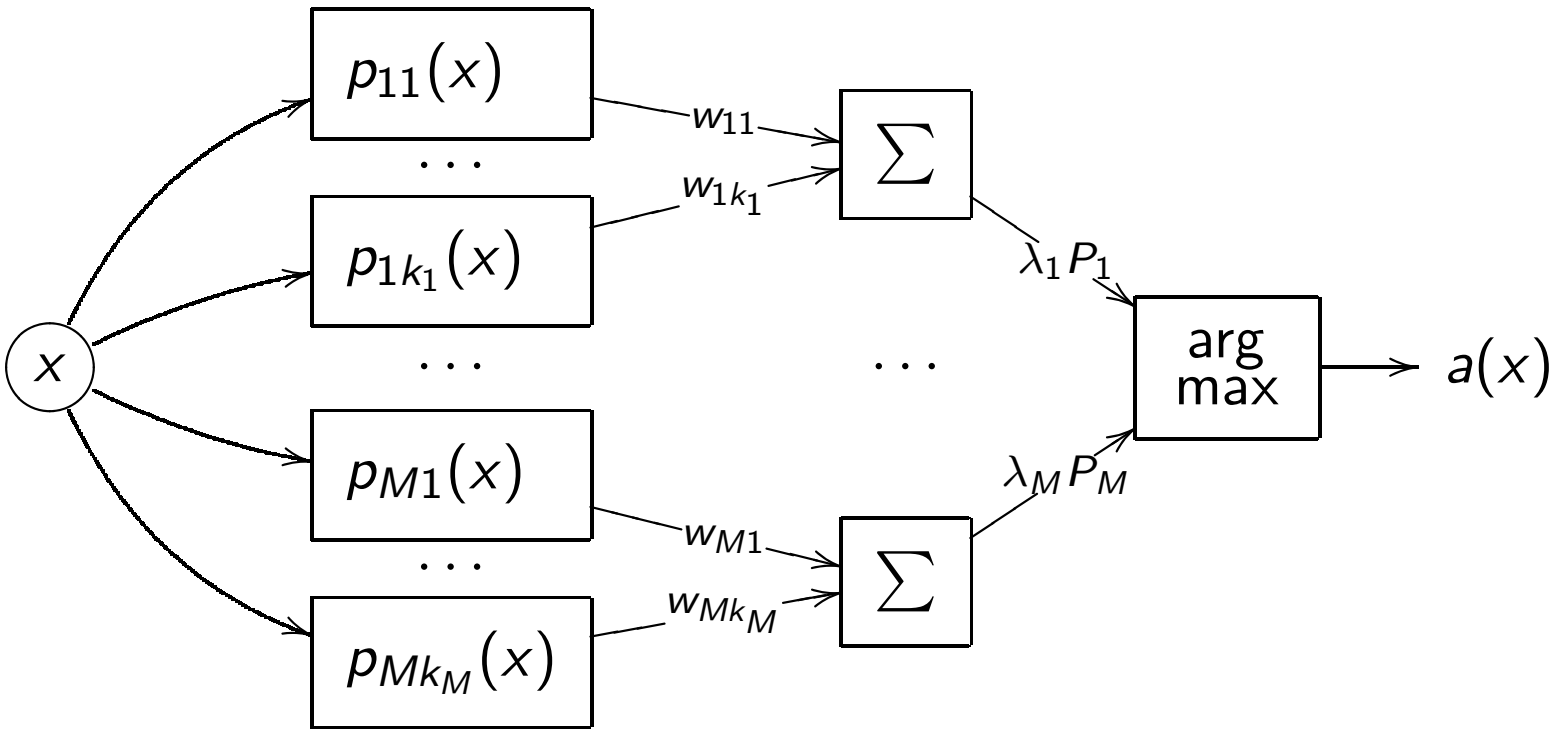
$$a(x) = \arg\max_{y \in Y} \lambda_y P_y \sum_{j=1}^{k_y} w_{yj} \underbrace{\mathcal{N}_{yj} \exp\left(-\tfrac{1}{2}\rho_{yj}^2(x, \mu_{yj})\right)}_{p_{yj}(x)},$$

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

$$\rho_{yj}^2(x, \mu_{yj}) = \sum_{d=1}^{n} \frac{1}{\sigma_{yjd}^2}\left(f_d(x) - \mu_{yjd}\right)^2.$$

Сеть радиальных базисных функций

Radial Basis Functions (RBF) — трёхуровневая суперпозиция:



Преимущества EM-RBF

EM — один из лучших алгоритмов обучения радиальных сетей.

Преимущества EM-алгоритма:

1. EM-алгоритм легко сделать устойчивым к шуму
2. EM-алгоритм довольно быстро сходится
3. автоматически строится *структурное описание* каждого класса в виде совокупности компонент — *кластеров*

Недостатки EM-алгоритма:

1. EM-алгоритм чувствителен к начальному приближению

Резюме в конце лекции

- Восстановление смеси — наиболее мощный подход к оцениванию плотности распределения по выборке.
- EM алгоритм сводит сложную многоэкстремальную задачу к серии стандартных подзадач максимизации правдоподобия для отдельных компонент смеси.
- EM алгоритм — очень мощная штука.
  Он применяется не только для восстановления смесей.
- У него есть масса обобщений: GEM, SEM, HEM,...
- Предполагая, что компоненты смеси — гауссовские с диагональными матрицами ковариации, получили метод обучения радиальных базисных функций

Общее резюме по байесовским классификаторам

- Эту формулу надо помнить: $a(x) = \arg\max_{y \in Y} \lambda_y P_y p_y(x)$.
- Три основных подхода к восстановлению функций правдоподобия $p_y(x)$ по выборке: параметрический, непараметрический и смесь распределений.
- Наивный байесовский классификатор основан на «драконовском» предположении о независимости признаков. Как ни странно, иногда это работает.
- Непараметрический подход наиболее прост, но возникает проблема выбора метрики.
- Параметрический подход требует задания вида распределения. Для примера мы ограничились гауссовскими.
- Восстановление смеси — наиболее гибкий подход. В случае гауссовских распределений он приводит к сильному методу — RBF (радиальных базисных функций).