Байесовская теория классификации и методы восстановления плотности

Воронцов Константин Вячеславович
vokov@forecsys.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

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

Видеолекции: http://shad.yandex.ru/lectures

12 апреля 2018

Содержание

Постановка задачи

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

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

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

Временное допущение: пусть известна совместная плотность

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

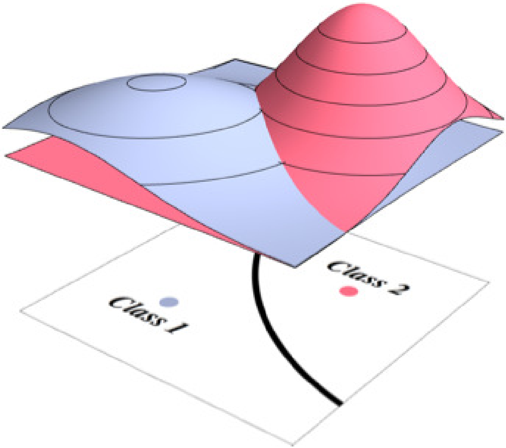
$\mathsf{P}(y)$ — *априорная вероятность* класса y;
$p(x|y)$ — *функция правдоподобия* класса y;
$\mathsf{P}(y|x)$ — *апостериорная вероятность* класса y;

Принцип максимума апостериорной вероятности:

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

Классификация по максимуму функции правдоподобия

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



Оптимальный байесовский классификатор

Теорема

Пусть $\mathrm{P}(y)$ 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 *байесовским классификатором*

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

Две подзадачи, причём вторая уже решена!

1. Восстановление плотности распределения по выборке
   Дано: $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка.
   Найти: эмпирические оценки $\hat{P}(y)$ и $\hat{p}(x|y)$, $y \in Y$
2. Построение классификатора
   Дано: вероятности $P(y)$ и плотности $p(x|y)$, $y \in Y$.
   Найти: классификатор $a\colon X \times Y$, минимизирующий $R(a)$.

Замечание 1: после замены $P(y)$ и $p(x|y)$ их эмпирическими оценками байесовский классификатор уже не оптимален.

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

Наивный байесовский классификатор

Допущение (действительно наивное):

Признаки $f_j\colon X \to D_j$ — 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:

$$p(x|y) = p_1(\xi_1|y) \cdots p_n(\xi_n|y), \quad x = (\xi_1, \dots, \xi_n), \quad y \in Y.$$

Прологарифмируем (для удобства). Получим классификатор

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

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

Восстановление одномерной плотности вероятности

Задача: по выборке $X^m = (x_i)_{i=1}^m$ оценить плотность $\hat{p}(x)$.

Дискретный случай: $|X| \ll m$. Гистограмма частот:

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

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 (заменяем вероятность на долю объектов выборки):

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

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:

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

где $K(r)$ — *ядро*, удовлетворяющее требованиям:
— чётная функция;
— нормированная функция: $\int K(r)\, dr = 1$;
— невозрастающая при $r > 0$, неотрицательная функция.

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

Метод парзеновского окна (Parzen window)

Многомерное обобщение: $\rho(x, x')$ — метрика на X.
Парзеновская оценка плотности для каждого класса $y \in Y$:

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

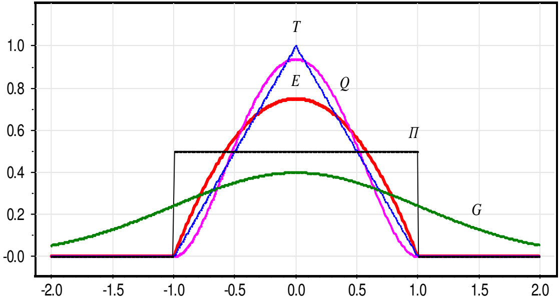
Метод окна Парзена — это метрический классификатор:

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

Замечание 1: нормирующий множитель
$V(h) = \int_X K\left(\frac{\rho(x,x_i)}{h}\right) dx$ не должен зависеть от x_i и y_i.

Замечание 2: имеем проблемы выбора
ядра $K(r)$, ширины окна h, функции расстояния $\rho(x, x')$.

Выбор ядра



$E(r) = \frac{3}{4}(1-r^2)\bigl[|r| \leqslant 1\bigr]$ — оптимальное (Епанечникова);

$Q(r) = \frac{15}{16}(1-r^2)^2\bigl[|r| \leqslant 1\bigr]$ — квартическое;

$T(r) = \bigl(1-|r|\bigr)\bigl[|r| \leqslant 1\bigr]$ — треугольное;

$G(r) = (2\pi)^{-1/2}\exp(-\frac{1}{2}r^2)$ — гауссовское;

$\Pi(r) = \frac{1}{2}\bigl[|r| \leqslant 1\bigr]$ — прямоугольное.

Выбор ядра почти не влияет на качество восстановления

Функционал качества восстановления плотности:

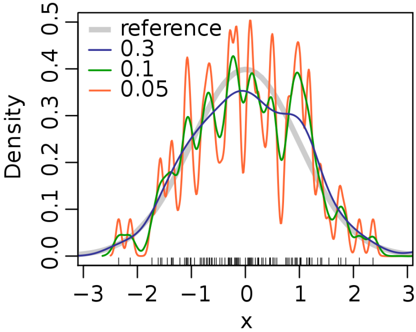
$$J(K) = \int_{-\infty}^{+\infty} \mathsf{E}\big(\hat{p}_h(x) - p(x)\big)^2\,dx.$$

Асимптотические значения отношения $J(K^*)/J(K)$ при $m \to \infty$ не зависят от вида распределения $p(x)$.

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

Пример. Зависимость оценки плотности от ширины окна

Оценка $\hat{p}_h(x)$ при различных значениях ширины окна h:



Вывод: Качество восстановления плотности существенно зависит от ширины окна h, но слабо зависит от вида ядра K.

Выбор ширины окна

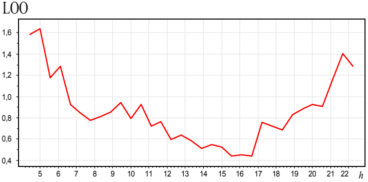
Скользящий контроль *Leave One Out* для классификации:

$$\mathrm{LOO}(h) = \sum_{i=1}^{\ell} \Bigl[a\bigl(x_i; X^\ell \backslash x_i, h\bigr) \neq y_i\Bigr] \to \min_h,$$

Leave One Out для восстановления плотности:

$$\mathrm{LOO}(h) = -\sum_{i=1}^{m} \ln \hat{p}_h(x_i; X^\ell \backslash x_i) \to \min_h,$$

Типичный вид зависимости $\mathrm{LOO}(h)$:



Окна переменной ширины

Проблема:

при наличии локальных сгущений любая h не оптимальна.

Идея:

задавать не ширину окна h, а число соседей k.

$$h_k(x) = \rho\bigl(x, x^{(k+1)}\bigr),$$

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

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

Замечание 1: нормировка $V(h_k)$ не должна зависеть от y, поэтому выборка ранжируется целиком, а не по классам X_y.

Замечание 2: оптимизация $\mathrm{LOO}(k)$ аналогична $\mathrm{LOO}(h)$.

Принцип максимума правдоподобия

Задана параметрическая модель плотности

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

где θ — параметр, φ — фиксированная функция.
Найдём оптимальное θ по i.i.d. выборке $X^m = \{x_1,\dots,x_m\}$.

Принцип максимума правдоподобия:

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

Необходимое условие оптимума:

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

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:

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

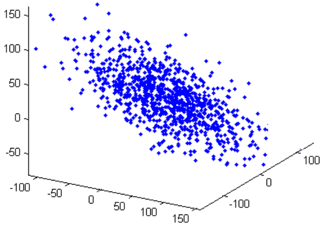
где $\mu_y \in \mathbb{R}^n$ — вектор матожидания (центр) класса $y \in Y$,
$\Sigma_y \in \mathbb{R}^{n\times n}$ — ковариационная матрица класса $y \in Y$ (симметричная, невырожденная, положительно определённая).

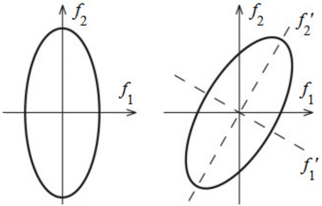
Теорема

1. Разделяющая поверхность
   $\bigl\{x \in X \bigm| \lambda_y P(y)p(x|y) = \lambda_s P(s)p(x|s)\bigr\}$
   квадратична для всех $y, s \in Y$, $y \neq s$.
2. Если $\Sigma_y = \Sigma_s$, то она вырождается в линейную.

Геометрический смысл предположения о нормальности классов

Каждый класс — облако точек эллиптической формы:





Если $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$: оси эллипсоида параллельны ортам

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

$V = (v_1, \ldots, v_n)$ — ортогональные собственные векторы Σ,

$S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ — собственные значения

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

$x' = V^{\mathsf{T}}x$ — *декоррелирующее* ортогональное преобразование

Квадратичный дискриминант

Теорема

Оценки максимального правдоподобия для n-мерных гауссовских плотностей классов $y \in Y$:

$$\hat{\mu}_y = \frac{1}{\ell_y} \sum_{i\colon y_i = y} x_i;$$

$$\hat{\Sigma}_y = \frac{1}{\ell_y} \sum_{i\colon y_i = y} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^{\mathsf{T}}.$$

Квадратичный дискриминант — подстановочный алгоритм:

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

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

Линейный дискриминант Фишера

Допущение:

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

Оценка максимума правдоподобия для Σ:

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

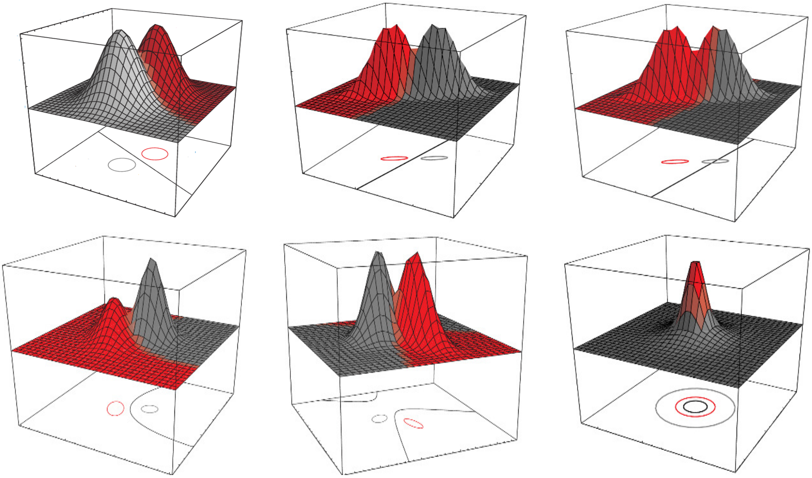
Линейный дискриминант — подстановочный алгоритм:

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

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

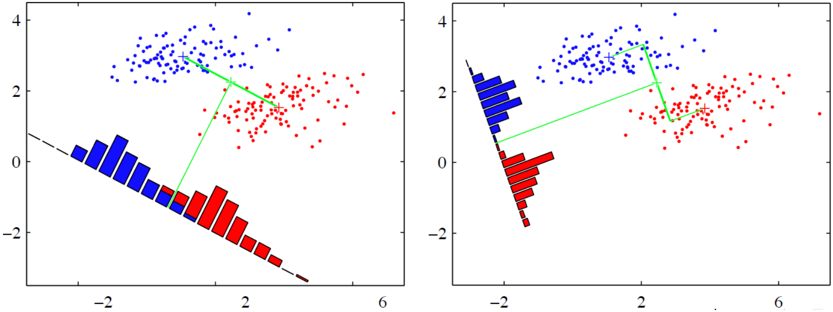
$$a(x) = \arg\max_{y\in Y}\ \bigl(x^{\mathsf{T}}\alpha_y + \beta_y\bigr).$$

Геометрический смысл квадратичного дискриминанта



Геометрический смысл линейного дискриминанта

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



Проблема мулитиколлинеарности

Матрица $\hat{\Sigma}_y$ вырождена при $\ell_y < n$
и может быть плохо обусловлена при $\ell_y \geqslant n$

- Регуляризация ковариационной матрицы:
  1) обращение $\hat{\Sigma} + \tau I_n$ вместо $\hat{\Sigma}$
  2) выбор параметра τ по скользящему контролю
- Диагонализация ковариационной матрицы,
  нормальный наивный байесовский классификатор:

$$a(x) = \arg\max_{y\in Y}\Bigl(\ln \lambda_y \hat{P}(y) + \sum_{j=1}^{n} \ln \hat{p}_j(\xi_j|y)\Bigr), \quad x \equiv (\xi_1, \dots, \xi_n);$$

$$\hat{p}_j(\xi|y) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_{yj}} \exp\Bigl(-\frac{(\xi-\hat{\mu}_{yj})^2}{2\hat{\sigma}_{yj}^2}\Bigr), \quad y \in Y, \quad j = 1, \dots, n;$$

$\hat{\mu}_{yj}$ и $\hat{\sigma}_{yj}$ — оценки среднего и дисперсии признака j в X_y.

Задача восстановления смеси распределений

Порождающая модель смеси распределений:

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

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

Задача 1: при фиксированном k,
имея простую выборку $X^m = \{x_1, \dots, x_m\} \sim p(x)$,
оценить вектор параметров $(w, \theta) = (w_1, \dots, w_k, \theta_1, \dots, \theta_k)$.

Задача 2: оценить ещё и k.

Максимизация правдоподобия и EM-алгоритм

Задача максимизации логарифма правдоподобия

$$L(w,\theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j \varphi(x_i,\theta_j) \to \max_{w,\theta}.$$

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

Итерационный алгоритм Expectation–Maximization:

1: начальное приближение параметров (w,θ);
2: **повторять**
3: оценка скрытых переменных $G = (g_{ij})$, $g_{ij} = P(j|x_i)$:
$G :=$ E-шаг (w,θ);
4: максимизация правдоподобия, отдельно по компонентам:
$(w,\theta) :=$ M-шаг (w,θ,G);
5: **пока** w,θ и G не стабилизируются.

EM-алгоритм как способ решения системы уравнений

Теорема (необходимые условия экстремума)

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

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

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

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

EM-алгоритм — это метод простых итераций для её решения

Вероятностная интерпретация

E-шаг — это формула Байеса:

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

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

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

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

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

Доказательство. Условия Каруша–Куна–Таккера

Лагранжиан оптимизационной задачи «$L(w,\theta) \to \max$»:

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

Приравниваем нулю производные:

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

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

■

EM-алгоритм

Вход: $X^m = \{x_1, \dots, x_m\}$, k, δ, начальные $(w_j, \theta_j)_{j=1}^k$;

Выход: $(w_j, \theta_j)_{j=1}^k$ — параметры смеси распределений

1: **повторять**

2: Е-шаг (expectation):

для всех $i = 1, \dots, m$, $j = 1, \dots, k$

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

3: М-шаг (maximization):

для всех $j = 1, \dots, k$

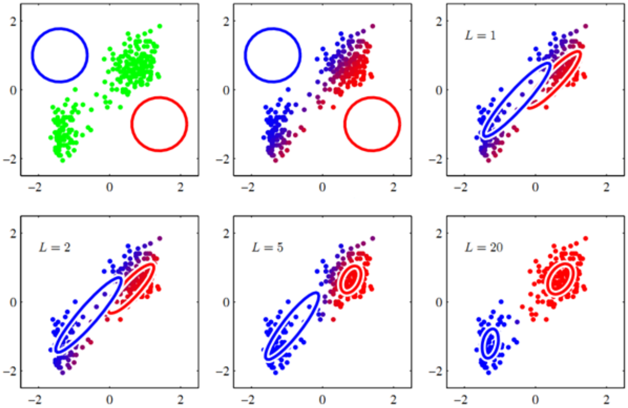
$$\theta_j := \arg\max_\theta \sum_{i=1}^m g_{ij} \ln \varphi(x_i, \theta); \qquad w_j := \frac{1}{m} \sum_{i=1}^m g_{ij};$$

4: **пока** $\max\limits_{i,j} |g_{ij} - g_{ij}^0| > \delta$;

5: **вернуть** $(w_j, \theta_j)_{j=1}^k$;

Пример

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



EM-алгоритм с добавлением и удалением компонент

Проблемы базового варианта EM-алгоритма:

- Как выбирать начальное приближение?
- Как определять число компонент?
- Как ускорить сходимость?

Добавление и удаление компонент в EM-алгоритме:

- Если слишком много объектов x_i имеют слишком низкие правдоподобия $p(x_i)$, то создаём новую $k{+}1$-ю компоненту, по этим объектам строим её начальное приближение.
- Если у j-й компоненты слишком низкий w_j, удаляем её.

  Регуляризация $L(w,\theta) - \tau \sum\limits_{j=1}^{k} \ln w_j \to \max$:

$$w_j \propto \left(\frac{1}{m} \sum_{i=1}^{m} g_{ij} - \tau \right)_+$$

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, $y \in Y = \{1, \dots, M\}$.

2. Компоненты имеют n-мерные гауссовские плотности с некоррелированными признаками:

$\mu_{yj} = (\mu_{yj1}, \dots, \mu_{yjn})$, $\Sigma_{yj} = \mathrm{diag}(\sigma^2_{yj1}, \dots, \sigma^2_{yjn})$, $j = 1, \dots, k_y$:

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

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

Эмпирические оценки средних и дисперсий

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

Решение задачи М-шага:
для всех классов $y \in Y$ и всех компонент $j = 1, \dots, k_y$,

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

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

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

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

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 $X = \mathbb{R}^n$:

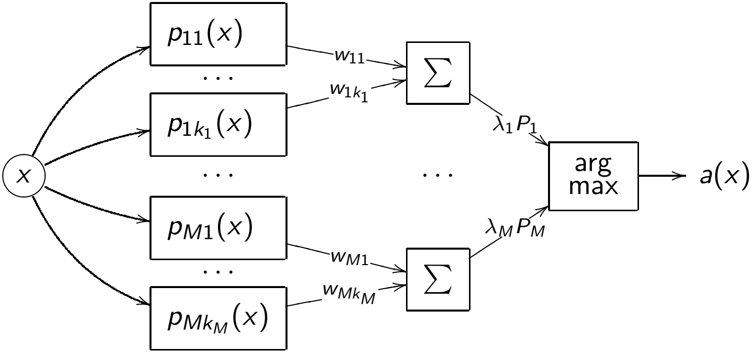
$$\rho^2_{yj}(x,\mu_{yj}) = \sum_{d=1}^{n} \frac{1}{\sigma^2_{yjd}}\bigl(f_d(x) - \mu_{yjd}\bigr)^2.$$

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

Байесовский классификатор — сеть RBF

Radial Basis Functions (RBF) — трёхуровневая суперпозиция:

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



Преимущества RBF-EM

EM — один из лучших алгоритмов обучения радиальных сетей.

Преимущества EM-алгоритма по сравнению с SVM:

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
2. Определение числа компонент — трудная задача (простые эвристики могут плохо работать)

Резюме

- Эту формулу полезно помнить:
  $a(x) = \arg\max_{y \in Y} \lambda_y P(y) p(x|y)$.
- Наивный байесовский классификатор:
  предположение о независимости признаков.
  Как ни странно, иногда это работает.
- Три подхода к восстановлению плотности $p(x|y)$ по выборке:
  - *Параметрический подход:*
    модель плотности + максимизация правдоподобия;
  - *Непараметрический подход:*
    наиболее прост и приводит к методу парзеновского окна;
  - *Разделение смеси распределений:*
    в случае смеси гауссиан приводит к методу RBF.