Байесовская теория классификации. Логистическая регрессия. Восстановление смеси плотностей.

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

Содержание

Напоминание. Байесовская теория классификации

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

Две подзадачи:

1. **Дано:**
   $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка.
   Найти:
   эмпирические оценки $\hat{P}(y)$ и $\hat{p}(x|y)$, $y \in Y$
   (восстановить плотность каждого класса по выборке).
2. **Дано:**
   априорные вероятности $P(y)$ и плотности $p(x|y)$, $y \in Y$.
   Найти:
   классификатор $a\colon X \times Y$, минимизирующий риск $R(a)$.
   Решение:
   $a(x) = \arg\max\limits_{y \in Y} \lambda_y P(y) p(x|y)$.

Логистическая регрессия: базовые предположения

- $X = \mathbb{R}^n$, $Y = \pm 1$, выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$ i.i.d. из

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

- Функции правдоподобия из *экспоненциального семейства*:

$$p(x|y) = \exp\bigl(c_y(\delta)\langle\theta_y, x\rangle + b_y(\delta, \theta_y) + d(x, \delta)\bigr),$$

где $\theta_y \in \mathbb{R}^n$ — параметр *сдвига*;
δ — параметр *разброса*;
b_y, c_y, d — произвольные числовые функции;
причём параметры $d(\cdot)$ и δ не зависят от y.

Экспоненциальное семейство распределений широко:
равномерное, нормальное, Лапласа, Пуассона, Парето, Дирихле, биномиальное, Г-распределение, χ^2-распределение, и др.

Пример: многомерное нормальное распределение

Многомерное нормальное распределение, $\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n}$,

$$\mathcal{N}(x;\mu,\Sigma) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} \exp\Bigl(-\tfrac{1}{2}(x-\mu)^\intercal\Sigma^{-1}(x-\mu)\Bigr)$$

принадлежит экспоненциальному семейству,
имеет параметры сдвига $\theta = \Sigma^{-1}\mu$ и разброса $\delta = \Sigma$:

$$\mathcal{N}(x;\mu,\Sigma) = \exp\Bigl(\underbrace{\mu^\intercal\Sigma^{-1}x}_{\langle\theta,x\rangle} - \underbrace{\tfrac{1}{2}\mu^\intercal\Sigma^{-1}\Sigma\Sigma^{-1}\mu}_{b(\delta,\theta)} - \underbrace{\tfrac{1}{2}x^\intercal\Sigma^{-1}x - \tfrac{n}{2}\ln(2\pi) - \tfrac{1}{2}\ln|\Sigma|}_{d(x,\delta)}\Bigr).$$

Нормальный байесовский классификатор линеен, если $\Sigma_y \equiv \Sigma$.
Но, может, класс плотностей, для которых он линеен, шире?

Теорема о линейности байесовского классификатора

Оптимальный байесовский классификатор для двух классов:

$$a(x) = \operatorname{sign}\bigl(\lambda_+ \,\mathsf{P}(+1|x) - \lambda_- \,\mathsf{P}(-1|x)\bigr) = \operatorname{sign}\left(\frac{p(x|+1)P(+1)}{p(x|-1)P(-1)} - \frac{\lambda_-}{\lambda_+}\right).$$

Теорема

Если $p(x|y)$ принадлежат экспоненциальному семейству, параметры $d(\cdot)$ и δ не зависят от y, и среди признаков $f_1(x), \ldots, f_n(x)$ есть константа, то байесовский классификатор линеен:

$$a(x) = \operatorname{sign}\bigl(\langle w, x\rangle - w_0\bigr), \qquad w_0 = \ln(\lambda_-/\lambda_+);$$

апостериорные вероятности классов:

$$\mathsf{P}(y|x) = \sigma\bigl(\langle w, x\rangle y\bigr),$$

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

Доказательство: шаг 1

После подстановки экспоненциальных плотностей классов

$$p(x|\pm 1) = \exp\bigl(c_\pm(\delta)\langle\theta_\pm, x\rangle + b_\pm(\delta,\theta_\pm) + d(x,\delta)\bigr)$$

в формулу байесовского классификатора

$$a(x) = \operatorname{sign}\left(\frac{\mathsf{P}(+1|x)}{\mathsf{P}(-1|x)} - \frac{\lambda_-}{\lambda_+}\right) = \operatorname{sign}\left(\ln\frac{P(+1)p(x|+1)}{P(-1)p(x|-1)} - \ln\frac{\lambda_-}{\lambda_+}\right)$$

получаем

$$\ln\frac{\mathsf{P}(+1|x)}{\mathsf{P}(-1|x)} = \langle\underbrace{c(\delta)(\theta_+ - \theta_-)}_{w=\mathrm{const}(x)}, x\rangle + \underbrace{b_+(\delta,\theta_+) - b_-(\delta,\theta_-) + \ln\tfrac{P_+}{P_-}}_{\beta=\mathrm{const}(x)}.$$

Добавим β к коэффициенту w_j при константном признаке $f_j = 1$

Основная теорема. Доказательство: шаг 2

Таким образом,

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

По формуле полной вероятности $\mathsf{P}(-1|x) + \mathsf{P}(+1|x) = 1$,

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

Объединяя эти два равенства в одно $(y = \pm 1)$, получаем:

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

Следовательно, разделяющая поверхность линейна:

$$\lambda_- \mathsf{P}(-1|x) = \lambda_+ \mathsf{P}(+1|x),$$

$$\langle w,x\rangle - \ln\frac{\lambda_-}{\lambda_+} = 0. \qquad \blacksquare$$

Обоснование логарифмической функции потерь

Максимизация логарифма правдоподобия выборки:

$$L(w) = \log \prod_{i=1}^{\ell} p(x_i, y_i) \to \max_w .$$

Подставим: $p(x, y) = \mathsf{P}(y|x) \cdot p(x) = \sigma\bigl(\langle w, x\rangle y\bigr) \cdot \mathrm{const}(w)$

$$L(w) = \sum_{i=1}^{\ell} \log \sigma\bigl(\langle w, x_i\rangle y_i\bigr) + \mathrm{const}(w) \to \max_w .$$

Максимизация $L(w)$ эквивалентна минимизации $Q(w)$:

$$Q(w) = \sum_{i=1}^{\ell} \log\bigl(1 + \exp(-\underbrace{\langle w, x_i\rangle y_i}_{M_i(w)})\bigr) \to \min_w .$$

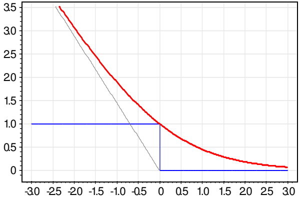
Задача классификации. Логистическая регрессия (LR)

$Y = \{-1, +1\}$ — два класса, $a(x, w) = \mathrm{sign}(\langle w, x \rangle)$, $x, w \in \mathbb{R}^n$.

Функционал аппроксимированного эмпирического риска:

$$Q(w) = \sum_{i=1}^{\ell} \bigl[M_i(w) < 0\bigr] \leqslant \sum_{i=1}^{\ell} \mathscr{L}\bigl(\langle w, x_i \rangle y_i\bigr) \to \min_w,$$

где $\mathscr{L}(M) = \log\bigl(1 + e^{-M}\bigr)$ — логарифмическая функция потерь



$M_i = \langle w, x_i \rangle y_i$

Напоминания. Оптимизация параметров LR.

- Метод первого порядка — стохастический градиент:

$$w^{(t+1)} := w^{(t)} + \eta_t \, y_i x_i \, (1 - \sigma_i),$$

η_t — градиентный шаг,
$\sigma_i = \sigma(y_i w^{\mathsf{T}} x_i) = \mathsf{P}(y_i | x_i)$ — вероятность правильной классификации x_i.

- Метод второго порядка (Ньютона-Рафсона) приводит к IRLS, Iteratively Reweighted Least Squares:

$$w^{(t+1)} := w^{(t)} + \eta_t (F^{\mathsf{T}} \Lambda F)^{-1} F^{\mathsf{T}} \tilde{y},$$

F — матрица объекты–признаки $\ell \times n$,
$\tilde{y} = \bigl(y_i(1 - \sigma_i)\bigr)$,
$\Lambda = \operatorname{diag}\bigl((1 - \sigma_i)/\sigma_i\bigr)$,

Пример. Бинаризация признаков и скоринговая карта

Задача кредитного скоринга:

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

Бинаризация признаков:

$$b_{jk}(x) = \big[f_j(x) \in D_{jk}\big]$$

$b_{jk}(x)$ — биномиальные с.в., из exp-семейства (многомерное распределение Бернулли)

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

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

$$R(x) = \sum_{y \in Y} D_{xy} \, \mathsf{P}(y|x) = \sum_{y \in Y} D_{xy} \sigma\bigl(\langle w, x \rangle y\bigr),$$

где D_{xy} — величина потери для (x, y).

Методика VaR (Value at Risk)

Оценивается не ожидаемая потеря, а распределение потерь:

- для каждого x_i разыгрывается N раз исход $y_i \sim P(y|x_i)$;
- строится эмпирическое распределение потерь $V = \sum\limits_{i=1}^{\ell} D_{x_i y_i}$;
- 99%-квантиль эмпирического распределения определяет величину резервируемого капитала

Задача восстановления смеси распределений

Порождающая модель смеси распределений:

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

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

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

Задача 2: оценить ещё и k.

Максимизация правдоподобия и EM-алгоритм

Задача максимизации логарифма правдоподобия

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

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

Проблема: задача не решается аналитически «в лоб».

Итерационный алгоритм Expectation–Maximization:

1: начальное приближение вектора параметров Θ;
2: **повторять**
3: $\quad G :=$ Е-шаг (Θ); // оцениваются *скрытые переменные* G
4: $\quad \Theta :=$ М-шаг (Θ, G);
5: **пока** Θ и G не стабилизируются.

EM-алгоритм как способ решения системы уравнений

Теорема (необходимые условия экстремума)

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

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

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

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

EM-алгоритм — это метод простых итераций для её решения

Вероятностная интерпретация

E-шаг — это формула Байеса:

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

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:

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

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

Доказательство. Условия Каруша–Куна–Таккера

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

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

Приравниваем нулю производные:

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

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

■

EM-алгоритм

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

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

1: **повторять**

2: E-шаг (expectation):

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

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

3: M-шаг (maximization):

для всех $j = 1, \dots, k$

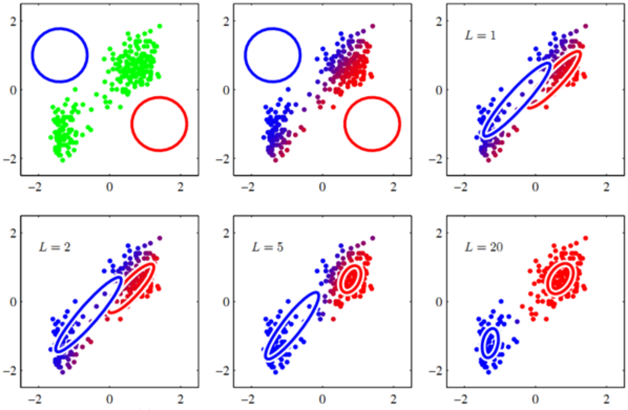
$$\theta_j := \arg\max_\theta \sum_{i=1}^m g_{ij} \ln p_j(x_i; \theta); \qquad w_j := \frac{1}{m} \sum_{i=1}^m g_{ij};$$

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:



ЕМ-алгоритм с добавлением и удалением компонент

Проблемы базового варианта ЕМ-алгоритма:

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

Добавление и удаление компонент в ЕМ-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 $L(\Theta) - \tau \sum\limits_{j=1}^{k} \ln w_j \to \max$:

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

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.

Преимущества:

- сохраняется свойство слабой локальной сходимости (в смысле увеличения правдоподобия на каждом шаге)
- повышается скорость сходимости при сопоставимом качестве решения

SEM — стохастический EM-алгоритм

Идея: на M-шаге вместо максимизации

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

максимизируется обычное, невзвешенное, правдоподобие

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

выборки X_j строятся путём стохастического моделирования:
для каждого $i = 1, \ldots, m$ генерируется $j \sim \mathsf{P}(\theta_j | x_i) \equiv g_{ij}$
и объект x_i помещается в X_j.

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

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

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

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

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

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

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

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

Алгоритм классификации

Подставим гауссовскую смесь в байесовский классификатор:

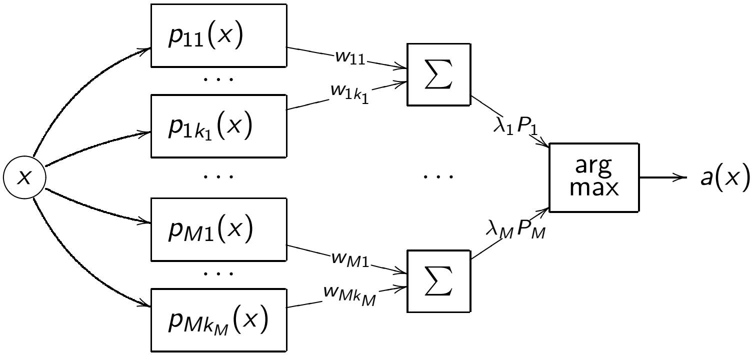
$$a(x) = \arg\max_{y \in Y} \lambda_y P_y \sum_{j=1}^{k_y} w_{yj} \underbrace{\mathcal{N}_{yj} \exp\left(-\tfrac{1}{2}\rho_{yj}^2(x, \mu_{yj})\right)}_{p_{yj}(x)},$$

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

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

Сеть радиальных базисных функций

Radial Basis Functions (RBF) — трёхуровневая суперпозиция:



Преимущества EM-RBF

EM — один из лучших алгоритмов обучения радиальных сетей.

Преимущества EM-алгоритма (перед SVM, ANN):

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)