

Логистическая регрессия Восстановление смеси плотностей.

К. В. Воронцов, А.В. Зухба
vokov@forecsys.ru
a_l@mail.ru

май 2016

- 1 **Логистическая регрессия**
 - Экспонентные семейства плотностей
 - Обоснование логистической регрессии
 - Задача кредитного скоринга

- 2 **Восстановление смеси распределений**
 - EM-алгоритм
 - Некоторые модификации EM-алгоритма
 - Сеть радиальных базисных функций

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 $a: X \times Y$, минимизирующий риск $R(a)$.

Решение:

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

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*:

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

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, Γ -распределение, χ^2 -распределение, и др.

Пример: многомерное нормальное распределение

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

$$\mathcal{N}(x; \mu, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

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

$$\begin{aligned} \mathcal{N}(x; \mu, \Sigma) = \exp\left(\underbrace{\mu^\top \Sigma^{-1} x}_{\langle \theta, x \rangle} - \underbrace{\frac{1}{2} \mu^\top \Sigma^{-1} \Sigma \Sigma^{-1} \mu}_{b(\delta, \theta)} - \right. \\ \left. \underbrace{\frac{1}{2} x^\top \Sigma^{-1} x - \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma|}_{d(x, \delta)} \right). \end{aligned}$$

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:

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

Теорема

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

$$a(x) = \text{sign}(\langle w, x \rangle - w_0), \quad w_0 = \ln(\lambda_- / \lambda_+);$$

апостериорные вероятности классов:

$$P(y|x) = \sigma(\langle w, x \rangle y),$$

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

Доказательство: шаг 1

После подстановки экспоненциальных плотностей классов

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

в формулу байесовского классификатора

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

получаем

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

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

Основная теорема. Доказательство: шаг 2

Таким образом,

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

По формуле полной вероятности $P(-1|x) + P(+1|x) = 1$,

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

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

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

Следовательно, разделяющая поверхность линейна:

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

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



Обоснование логарифмической функции потерь

Максимизация логарифма правдоподобия выборки:

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

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

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

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

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

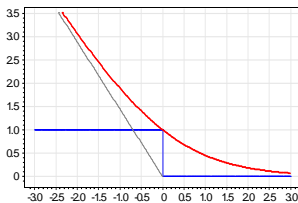
Задача классификации. Логистическая регрессия (LR)

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

Функционал аппроксимированного эмпирического риска:

$$Q(w) = \sum_{i=1}^{\ell} [M_i(w) < 0] \leq \sum_{i=1}^{\ell} \mathcal{L}(\langle w, x_i \rangle y_i) \rightarrow \min_w,$$

где $\mathcal{L}(M) = \log(1 + e^{-M})$ — 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,

$\sigma_i = \sigma(y_i w^\top x_i) = P(y_i | x_i)$ — вероятность правильной классификации x_i .

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

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

F — матрица объекты–признаки $\ell \times n$,

$\tilde{y} = (y_i(1 - \sigma_i))$,

$\Lambda = \text{diag}((1 - \sigma_i)/\sigma_i)$,

Пример. Бинаризация признаков и скоринговая карта

Задача кредитного скоринга:

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

Бинаризация признаков $f_j(x)$:

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

$b_{jk}(x)$ — биномиальные с.в.,
из эксп-семейства

(многомерное распределение

Бернулли тоже из эксп-семейства)

| | | |
|-------------------|-------------------------|----|
| 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 :

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

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 $V = \sum_{i=1}^{\ell} D_{x_i y_i}$;
- 99%-квантиль эмпирического распределения определяет величину резервируемого капитала

Задача восстановления смеси распределений

Порождающая модель смеси распределений:

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

$p_j(x; \theta_j)$ — 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 $\Theta = (w_1, \dots, w_k, \theta_1, \dots, \theta_k)$.

Задача 2: оценить ещё и k .

Максимизация правдоподобия и ЕМ-алгоритм

Задача максимизации логарифма правдоподобия

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

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

Проблема: задача не решается аналитически «в лоб».

Итерационный алгоритм Expectation–Maximization:

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

ЕМ-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 $G = (g_{ij})$:

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

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

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

ЕМ-алгоритм — это метод простых итераций для её решения

Вероятностная интерпретация

Е-шаг — это формула Байеса:

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

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

М-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 « $L(\Theta) \rightarrow \max$ »:

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

Приравниваем нулю производные:

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

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



ЕМ-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: Е-шаг (expectation):

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

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

3: М-шаг (maximization):

для всех $j = 1, \dots, k$

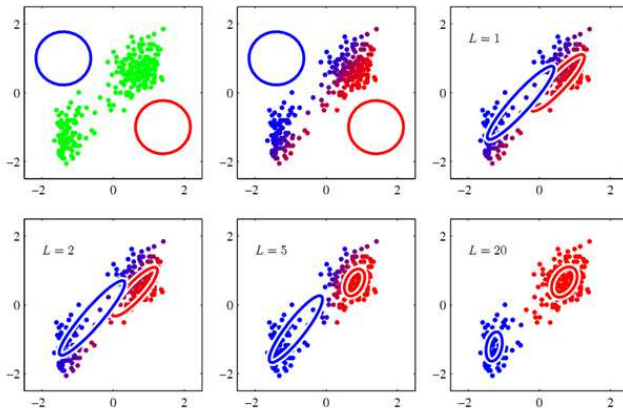
$$\theta_j := \arg \max_{\theta} \sum_{i=1}^m g_{ij} \ln p_j(x_i; \theta); \quad w_j := \frac{1}{m} \sum_{i=1}^m g_{ij};$$

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

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 $L(\Theta) - \tau \sum_{j=1}^k \ln w_j \rightarrow \max:$

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

GEM — обобщённый ЕМ-алгоритм

Идея:

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

Преимущества:

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

SEM — стохастический ЕМ-алгоритм

Идея: на М-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 $i = 1, \dots, m$ генерируется $j \sim P(\theta_j | x_i) \equiv g_{ij}$
и объект x_i помещается в X_j .

Преимущества:

ускорение сходимости, предотвращение закликиваний.

НЕМ — иерархический ЕМ-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:

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

$$p(x|y) = \sum_{j=1}^{k_y} w_{yj} p_{yj}(x), \quad p_{yj}(x) = \mathcal{N}(x; \mu_{yj}, \Sigma_{yj}),$$
$$\sum_{j=1}^{k_y} w_{yj} = 1, \quad w_{yj} \geq 0;$$

Эмпирические оценки средних и дисперсий

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

Решение задачи М-шага:

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

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

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

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

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

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

Алгоритм классификации

Подставим гауссовскую смесь в байесовский классификатор:

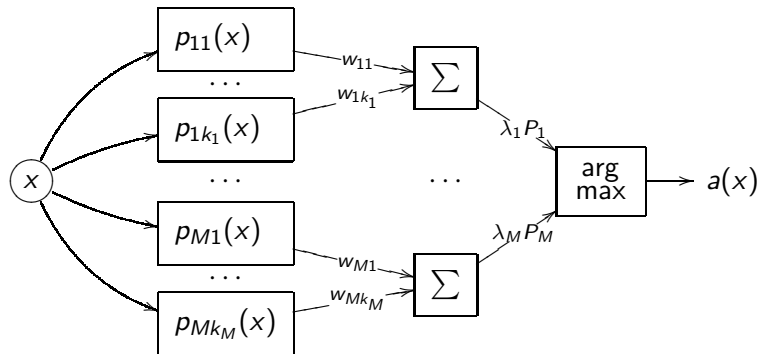
$$a(x) = \arg \max_{y \in Y} \lambda_y P_y \sum_{j=1}^{k_y} w_{yj} \underbrace{\mathcal{N}_{yj} \exp \left(-\frac{1}{2} \rho_{yj}^2(x, \mu_{yj}) \right)}_{\rho_{yj}(x)},$$

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

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

Сеть радиальных базисных функций

Radial Basis Functions (RBF) — трёхуровневая суперпозиция:



Преимущества ЕМ-RBF

ЕМ — один из лучших алгоритмов обучения радиальных сетей.

Преимущества ЕМ-алгоритма (перед SVM, ANN):

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

Недостатки ЕМ-алгоритма:

- 1 ЕМ-алгоритм чувствителен к начальному приближению
- 2 Определение числа компонент — трудная задача (простые эвристики могут плохо работать)