

# Семинары по байесовским методам

Евгений Соколов  
sokolov.evg@gmail.com

4 декабря 2015 г.

## 1 Байесовские методы машинного обучения

Пусть  $X = \{x_1, \dots, x_\ell\}$  — выборка,  $\mathbb{X}$  — множество всех возможных объектов,  $Y$  — множество ответов. В байесовском подходе предполагается, что обучающие объекты и ответы на них  $(x_1, y_1), \dots, (x_\ell, y_\ell)$  независимо выбираются из некоторого распределения  $p(x, y)$ , заданного на множестве  $\mathbb{X} \times Y$ . Данное распределение можно переписать как

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

где  $p(y)$  определяет вероятности появления каждого из возможных ответов и называется *априорным распределением*, а  $p(x | y)$  задает распределение объектов при фиксированном ответе  $y$  и называется *функцией правдоподобия*.

Если известны априорное распределение и функция правдоподобия, то по формуле Байеса можно записать *апостериорное распределение* на множестве ответов:

$$p(y | x) = \frac{p(x | y)p(y)}{\int_s p(x | s)p(s)ds} = \frac{p(x | y)p(y)}{p(x)},$$

где знаменатель не зависит от  $y$  и является нормировочной константой.

### §1.1 Оптимальные байесовские правила

Пусть на множестве всех пар ответов  $Y \times Y$  задана функция потерь  $L(y, s)$ . Наиболее распространенным примером для задач классификации является ошибка классификации  $L(y, s) = [y \neq s]$ , для задач регрессии — квадратичная функция потерь  $L(y, x) = (y - s)^2$ . *Функционалом среднего риска* называется матожидание функции потерь по всем парам  $(x, y)$  при использовании алгоритма  $a(x)$ :

$$R(a) = \mathbb{E}L(y, a(x)) = \int_Y \int_{\mathbb{X}} L(y, a(x))p(x, y)dx dy.$$

Если распределение  $p(x, y)$  известно, то можно найти алгоритм  $a_*(x)$ , оптимальный с точки зрения функционала среднего риска.

### 1.1.1 Классификация

Начнем с задачи классификации с множеством ответом  $Y = \{1, \dots, K\}$  и функции потерь  $L(y, s) = [y \neq s]$ . Покажем, что минимум функционала среднего риска достигается на алгоритме

$$a_*(x) = \arg \max_{y \in Y} p(y | x).$$

Для произвольного классификатора  $a(x)$  выполнена следующая цепочка неравенств [1]:

$$\begin{aligned} R(a) &= \int_Y \int_{\mathbb{X}} L(y, a(x)) p(x, y) dx dy = \\ &= \sum_{y=1}^K \int_{\mathbb{X}} [y \neq a(x)] p(x, y) dx = \\ &= \int_{\mathbb{X}} \sum_{y \neq a(x)} p(x, y) dx = \left\{ \int_{\mathbb{X}} \sum_{y \neq a(x)} p(x, y) dx + \int_{\mathbb{X}} p(x, a(x)) dx = 1 \right\} = \\ &= 1 - \int_{\mathbb{X}} p(x, a(x)) dx \geq \\ &\geq 1 - \int_{\mathbb{X}} \max_{s \in Y} p(x, s) dx = \\ &= 1 - \int_{\mathbb{X}} p(x, a_*(x)) dx = \\ &= R(a_*) \end{aligned}$$

Таким образом, средний риск любого классификатора  $a(x)$  не превосходит средний риск нашего классификатора  $a_*(x)$ .

Мы получили, что оптимальный байесовский классификатор выбирает тот класс, который имеет наибольшую апостериорную вероятность. Такой классификатор называется *MAP-классификатором* (maximum a posteriori).

### 1.1.2 Регрессия

Перейдем к задаче регрессии и функции потерь  $L(y, x) = (y - s)^2$ . Нам пригодится понятие условного матожидания:

$$\mathbb{E}(y | x) = \int_Y y p(y | x) dy.$$

Преобразуем функцию потерь [2]:

$$\begin{aligned} L(y, a(x)) &= (y - a(x))^2 = (y - \mathbb{E}(y | x) + \mathbb{E}(y | x) - a(x))^2 = \\ &= (y - \mathbb{E}(y | x))^2 + 2(y - \mathbb{E}(y | x))(\mathbb{E}(y | x) - a(x)) + (\mathbb{E}(y | x) - a(x))^2. \end{aligned}$$

Подставляя ее в функционал среднего риска, получаем:

$$\begin{aligned} R(a) &= \int_Y \int_{\mathbb{X}} L(y, a(x)) p(x, y) dx dy = \\ &= \int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t | x))^2 p(x, y) dx dy + \int_Y \int_{\mathbb{X}} (\mathbb{E}(t | x) - a(x))^2 p(x, y) dx dy + \\ &+ 2 \int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t | x)) (\mathbb{E}(t | x) - a(x)) p(x, y) dx dy. \end{aligned}$$

Разберемся сначала с последним слагаемым. Заметим, что величина  $(\mathbb{E}(t | x) - a(x))$  не зависит от  $y$ , и поэтому ее можно вынести за интеграл по  $y$ :

$$\begin{aligned} &\int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t | x)) (\mathbb{E}(t | x) - a(x)) p(x, y) dx dy = \\ &= \int_{\mathbb{X}} (\mathbb{E}(t | x) - a(x)) \int_Y \{(y - \mathbb{E}(t | x)) p(x, y)\} dy dx = \\ &= \int_{\mathbb{X}} (\mathbb{E}(t | x) - a(x)) \left\{ \int_Y y p(x, y) dy - \int_Y \mathbb{E}(t | x) p(x, y) dy \right\} dx = \\ &= \int_{\mathbb{X}} (\mathbb{E}(t | x) - a(x)) \left\{ p(x) \int_Y y p(y | x) dy - \mathbb{E}(t | x) \int_Y p(x, y) dy \right\} dx = \\ &= \int_{\mathbb{X}} (\mathbb{E}(t | x) - a(x)) \underbrace{\{p(x) \mathbb{E}(t | x) - p(x) \mathbb{E}(t | x)\}}_{=0} dx = \\ &= 0 \end{aligned}$$

Получаем, что функционал среднего риска имеет вид

$$R(a) = \int_Y \int_{\mathbb{X}} (y - \mathbb{E}(t | x))^2 p(x, y) dx dy + \int_Y \int_{\mathbb{X}} (\mathbb{E}(t | x) - a(x))^2 p(x, y) dx dy.$$

От алгоритма  $a(x)$  зависит только второе слагаемое, и оно достигает своего минимума, если  $a(x) = \mathbb{E}(t | x)$ . Таким образом, оптимальная байесовская функция регрессии для квадратичной функции потерь имеет вид

$$a_*(x) = \mathbb{E}(y | x) = \int_Y y p(y | x) dy.$$

Иными словами, мы должны провести «взвешенное голосование» по всем возможным ответам, причем вес ответа равен его апостериорной вероятности.

## §1.2 Байесовский вывод

Основной проблемой оптимальных байесовских алгоритмов, о которых шла речь в предыдущем разделе, является невозможность их построения на практике, поскольку нам никогда неизвестно распределение  $p(x, y)$ . Данное распределение можно попробовать восстановить по обучающей выборке, при этом существует два подхода — параметрический и непараметрический. Сейчас мы сосредоточимся на параметрическом подходе.

Допустим, распределение на парах «объект-ответ» зависит от некоторого параметра  $\theta$ :  $p(x, y | \theta)$ . Тогда получаем следующую формулу для апостериорной вероятности:

$$p(y | x, \theta) \propto p(x | y, \theta)p(y),$$

где выражение « $a \propto b$ » означает « $a$  пропорционально  $b$ ». Для оценивания параметров применяется *метод максимального правдоподобия*:

$$\theta_* = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \prod_{i=1}^{\ell} p(x_i | y_i, \theta),$$

где  $L(\theta)$  — функция правдоподобия. Примером такого подхода может служить *нормальный дискриминантный анализ*, где предполагается, что функции правдоподобия являются нормальными распределениями с неизвестными параметрами  $\theta = (\mu, \Sigma)$ . Об этом подходе речь пойдет на следующем семинаре, а сейчас рассмотрим более простой пример.

Иногда удобнее сразу задавать апостериорное распределение — например, в случае с линейной регрессией. Будем считать, что задан некоторый вектор весов  $w$ , и метка объекта  $y(x)$  генерируется следующим образом: вычисляется линейная функция  $\langle w, x \rangle$ , и к результату прибавляется нормальный шум:

$$y(x) = \langle w, x \rangle + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

В этом случае апостериорное распределение примет вид

$$p(y | x, w) = \mathcal{N}(\langle w, x \rangle, \sigma^2). \quad (1.1)$$

**Задача 1.1.** Покажите, что метод максимального правдоподобия для модели (1.1) эквивалентен методу наименьших квадратов.

**Решение.** Запишем правдоподобие для выборки  $x_1, \dots, x_\ell$ :

$$L(w) = \prod_{i=1}^{\ell} p(y_i | x_i, w) = \prod_{i=1}^{\ell} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_i - \langle w, x_i \rangle)^2}{2\sigma^2} \right).$$

Перейдем к логарифму правдоподобия:

$$\log L(w) = -\ell \log \sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{\ell} (y_i - \langle w, x_i \rangle)^2 \rightarrow \max_w.$$

Убирая все члены, не зависящие от вектора весов  $w$ , получаем задачу наименьших квадратов

$$\sum_{i=1}^{\ell} (y_i - \langle w, x_i \rangle)^2 \rightarrow \min_w.$$

■

**Байесовский вывод параметров.** В некоторых случаях применение метода максимального правдоподобия для поиска параметров приводит к плохим результатам. Например, если имеет место мультиколлинеарность, то функция правдоподобия имеет много минимумов, и решение может оказаться переобученным. Одним из подходов к устранению этой проблемы является введение априорного распределения *на параметрах*.

Пусть  $p(\theta)$  — априорное распределение на векторе параметров  $\theta$ . В качестве функции правдоподобия для данного вектора возьмем апостериорное распределение на ответах  $p(y | x, \theta)$ . Тогда по формуле Байеса

$$p(\theta | y, x) = \frac{p(y | x, \theta)p(\theta)}{p(y | x)}.$$

Вернемся к примеру с линейной регрессией. Введем априорное распределение на векторе весов:

$$p(w_j) = \mathcal{N}(0, \alpha^2), \quad j = 1, \dots, d.$$

Иными словами, мы предполагаем, что веса концентрируются вокруг нуля.

**Задача 1.2.** Покажите, что максимизация апостериорной вероятности  $p(w | y, x)$  для модели линейной регрессии с нормальным априорным распределением эквивалентна решению задачи гребневой регрессии.

**Решение.** Запишем апостериорную вероятность вектора весов  $w$  для выборки  $x_1, \dots, x_\ell$ :

$$\begin{aligned} p(w | y, x) &= \prod_{i=1}^{\ell} p(y_i | x_i, w) p(w) = \\ &= \prod_{i=1}^{\ell} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \langle w, x_i \rangle)^2}{2\sigma^2}\right) \prod_{j=1}^d \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{w_j^2}{2\alpha^2}\right). \end{aligned}$$

Перейдем к логарифму и избавимся от константных членов:

$$\log p(w | y, x) = -\frac{1}{2\sigma^2} \sum_{i=1}^{\ell} (y_i - \langle w, x_i \rangle)^2 - \underbrace{\frac{\ell}{2\alpha^2} \sum_{j=1}^d w_j^2}_{=\|w\|^2}.$$

В итоге получаем задачу гребневой регрессии

$$\sum_{i=1}^{\ell} (y_i - \langle w, x_i \rangle)^2 + \lambda \|w\|^2 \rightarrow \min_w,$$

где  $\lambda = \frac{\ell}{2\alpha^2}$ .

■

После того, как оптимальный вектор весов  $w_*$  найден, мы можем найти распределение на ответах для нового объекта  $x$ :

$$p(y | x, X, w_*) = \mathcal{N}(\langle x, w_* \rangle, \sigma^2).$$

Выше мы выяснили, что оптимальным ответом будет матожидание  $\mathbb{E}(y|x) = \int yp(y|x, X, w_*)dy$ .

С точки зрения байесовского подхода [3] правильнее не искать моду <sup>1</sup>  $w_*$  апостериорного распределения на параметрах и брать соответствующую ей модель  $p(y|x, X, w_*)$ , а устроить «взвешенное голосование» всех возможных моделей:

$$p(y|x, X) = \int p(y|x, w)p(w|Y, X)dw,$$

где  $X = \{x_1, \dots, x_\ell\}$ ,  $Y = \{y_1, \dots, y_\ell\}$ .

## Список литературы

- [1] *Ветров, Д.П., Кропотов, Д.А.* Байесовские методы машинного обучения. Учебное пособие. // Москва, 2007.
- [2] *Bishop, C.M.* Pattern Recognition and Machine Learning. // Springer, 2006.
- [3] *Murphy, K.P.* Machine Learning: A Probabilistic Perspective. // MIT Press, 2012.

---

<sup>1</sup> Мода — точка максимума плотности.