# Машинное обучение, ФКН ВШЭ
# Семинар №16

## 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

$$p(x, y) = p(y)p(x \mid y),$$

где $p(y)$ определяет вероятности появления каждого из возможных ответов и называется *априорным распределением*, а $p(x \mid y)$ 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:

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

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 $a(x)$:

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

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

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

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

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

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

Напомним, что при выводе разложения на шум, смещение и разброс функционала среднего риска для задачи регрессии и функции потерь $L(y, x) = (y - s)^2$ нами уже была получена формула оптимального алгоритма с точки зрения данного функционала:

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

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

## §1.2 Метод максимального правдоподобия

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 $\theta$: $p(x, y \mid \theta)$. Тогда получаем следующую формулу для апостериорной вероятности:

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

где выражение «$a \propto b$» означает «$a$ пропорционально $b$».

Если мы предполагаем, что вектор параметров $\theta$ константа, мы находимся в зоне действия частотной статистики. Оценить вектор параметров можно с помощью *метода максимального правдоподобия*:

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

где $L(\theta)$ — функция правдоподобия.

Примером такого подхода может служить *нормальный дискриминантный анализ*, где предполагается, что функции правдоподобия являются нормальными распределениями с неизвестными параметрами $\theta = (\mu, \Sigma)$.

Перед обучением разных моделей, мы всегда выбирали функцию потерь. Обычно мы делали это исходя из инженерных соображений. Для задачи классификации мы использовали логистические потери, logloss. Мы ввели их, как непрерывную функцию, ограничивающую сверху долю допущенный ошибок. Нам нужно было, чтобы функция потерь оказалась дифференцируемой.

Для линейной регрессии мы использовали квадратичные потери, MSE. Эта функция сильнее штрафует за большие ошибки и дифференцируема. Более того, её широкой применение можно обосновать следующим образом. Пусть:

1. Функция потерь представима в виде $L(y, a(x)) = g(y - a(x))$,
2. Если ответ верный, тогда ошибка нулевая, $g(0) = 0$,
3. Чем больше отклонение, тем выше ошибка $|z_1| \leq |z_2| \Rightarrow g(z_1) \leq g(z_2)$
4. У функции $g(z)$ существуют первые две производные,

тогда можно разложить функцию $g(z)$ в ряд Тэйлора в окрестности нуля

$$L(y, a(x)) = g(y - a(x)) \approx g(0) + g'(0) \cdot (y - a(x)) + \frac{g''(0)}{2} \cdot (y - a(x))^2 \approx C \cdot (y - a(x))^2.$$

Здесь мы воспользовались нашими предположениями 2 и 3. Таким образом, с точностью до константы, когда значения $y$ и $a(x)$ близки разумно использовать $MSE$.

Когда перед нами возникала какая-то проблема, например, выбросы. Мы пытались улучшить нашу функцию потерь. Так мы придумали MAE, функцию Хубера и log-cosh.

Вероятностный подход к машинному обучению предлагает альтернативный подход к получению функции потерь. Её можно вывести из метода максимального правдоподобия. Для того, чтобы этот приём сработал, нам нужно предположить как именно распределены данные.

Рассмотрим линейную регрессию. 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 распределение данных описывается как

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

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

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

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

Перейдем к логарифму правдоподобия:

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

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

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

При постановке задачи никто не требовал от нас оценки дисперсии. Поэтому мы её удалили из итоговой функции потерь. Также мы предположили, что дисперсия одинакова для всех наблюдений. Такое предположение называется *гомоскедастичностью*. На практике возникают задачи, в которых условная дисперсия $y_i$ внутри выборки может меняться. Это называют *гетероскедастичностью*.

Например, в случае недвижимости, разброс цен для элитной недвижимости может быть намного выше, чем разброс цен для квартир эконом-класса. При оценивании модели можно учесть это, взяв величины $\frac{1}{2\sigma_i^2}$ в нашу функцию потерь в качестве весов. Метод максимального правдоподобия позволяет оценить их вместе с параметрами $w$.

■

Снова рассмотрим линейную регрессию. В этот раз будем считать, что ошибка имеет распределение Лапласа. Ошибка в таком случае обладает плотностью распределения

$$p(t) = \frac{1}{2\sigma} \cdot e^{-\frac{|t|}{\sigma^2}}$$

**Задача 1.2.** *Покажите, что метод максимального правдоподобия в данном случае эквивалентен минимизации MAE.*

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

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

Перейдем к логарифму правдоподобия:

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

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

$$\sum_{i=1}^{\ell} |y_i - \langle w, x_i\rangle| \to \min_w.$$

■

Вероятностный подход позволяет по-новому взглянуть на недостатки функций потерь. Квадратичная ошибка чувствительная к выбросам, потому что мы предполагаем, что их нет. У нормально распределённых ошибок очень тонкие хвосты распределения. В случае распределения Лапласа, мы предполагаем толстые хвосты распределения ошибок. Поэтому MAE оказывается более устойчивой к выбросам.

Если мы знаем, какими свойствами обладают наши данные, мы можем заложить их в распределение и получить подходящую функцию потерь для оптимизации.

**Задача 1.3.** *Пусть переменная $y_i$ принимает только целочисленные значения. Например, это лайки на странице Маши в Instagram. Она получает их с какой-то интенсивностью $\lambda$, зависящей от характеристик её постов $x_i$. Например, может быть, что $\lambda = \lambda(x_i) = \langle w, x_i\rangle$. Такая модель называется пуассоновской регрессией. Какую функцию потерь нужно минимизировать, чтобы получить оценку $w$, исходя из принципа максимизации правдоподобия?*

**Решение.** Для распределения Пуассона

$$P(y = k) = \frac{e^{-\lambda}(\lambda)^k}{k!}$$

Выписываем функцию правдоподобия

$$L(w) = \prod_{i=1}^{\ell} \frac{e^{-\lambda(x_i)} \cdot (\lambda(x_i))^{y_i}}{y_i!} \to \max_w$$

Прологарифмируем

$$\ln L(w) = \sum_{i=1}^{\ell} y_i \log \lambda(x_i) - \lambda(x_i) - \log(y_i!) \to \max_w$$

Откидываем все константные слагаемые, домножаем на $-1$ и получаем функцию потерь для минимизации

$$\sum_{i=1}^{\ell}[\lambda(x_i) - y_i \log \lambda(x_i)] \to \min_w$$

В качестве $\lambda(x_i)$ можно использовать не только линейные модели. В библиотеке для градиентного бустинга catboost есть возможность использовать пуассоновскую регрессию с градиентным бустингом над деревьями[1].

■

Пуассоновская функция потерь окажется полезной для нас, когда мы будем говорить про неотрицательные матричные разложения в рекомендательных системах.

Метод максимального правдоподобия обладает хорошими асимптотическими свойствами. Если $\frac{\ell}{d} \to \infty$, где $\ell$ — число наблюдений, а $d$ — число оцениваемых параметров, тогда оценки максимального правдоподобия оказываются оптимальными с точки зрения их статистических свойств.

В современном машинном обучении величины $\ell$ и $d$ оказываются очень большими. Модели, с которыми мы работаем, оказываются перепараметризованными. Например, нейронные сети обычно обладают настолько большим числом параметров, что $\frac{\ell}{d} < 1$. В таких ситуациях становится довольно легко переобучиться.

**Задача 1.4.** *Пусть заданы выборка $X^\ell$ и распределение на объектах $p(x \mid \theta)$, параметр которого мы хотим настроить под данную выборку. Эмпирическим распределением называется дискретное распределение на объектах, присваивающее каждому объекту из обучающей выборки вероятность $1/\ell$:*

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

*Покажите, что максимизация правдоподобия эквивалентна минимизации дивергенции Кульбака-Лейблера между эмпирическим распределением и модельным распределением: $KL\left(\hat{p}(x \mid X^\ell) \,\middle\|\, p(x \mid \theta)\right)$.*

**Решение.** Распишем указанную дивергенцию:

$$\begin{aligned}\mathrm{KL}\left(\hat{p}(x \mid X^\ell) \,\middle\|\, p(x \mid \theta)\right) &= \sum_{i=1}^{\ell} \frac{1}{\ell} \log \frac{1/\ell}{p(x_i \mid \theta)} = \\ &= \sum_{i=1}^{\ell} \frac{1}{\ell} \log \frac{1}{\ell} - \frac{1}{\ell} \sum_{i=1}^{\ell} \log p(x_i \mid \theta) \to \min_\theta .\end{aligned}$$

Отбросим константные члены:

$$\sum_{i=1}^{\ell} \log p(x_i \mid \theta) \to \max_\theta .$$

[1] https://github.com/catboost/catboost/tree/master/catboost/tutorials/regression

Мы получили задачу максимизации логарифма правдоподобия.

Таким образом, метод максимума правдоподобия старается подобрать такие параметры модели, чтобы она давала равномерное распределение на объектах выборки и присваивала нулевую вероятность всем остальным объектам.

■

Чтобы защититься от переобучения в современном машинном обучении активно используют различные техники регуляризации. Сложная модель с большим числом параметров оказывается несмещенной, но обладает высоким разбросом. Регуляризация позволяет уменьшить разброс за счет некоторого смещения. Техники регуляризации можно обосновать с точки зрения байесовского вывода с помощью введения априорного распределения *на параметрах*.

## §1.3 Байесовский вывод

Если мы предполагаем, что вектор параметров $\theta$ случайная величина, мы находимся в зоне действия байесовской статистики. В байесовской статистике используются те же самые модели, что и в частотной, но способ оценивания параметров меняется с максимизации правдоподобия на байесовкий вывод.

Так как $\theta$ случайная величина, нам надо высказать своё мнение о ней в терминах плотностей распределения. Обычно такое мнение о распределении параметра называют априорным распределением.

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

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

После байесовского вывода мы получаем для каждого параметра в качестве оценки целое апостериорное распределение. С помощью него обычно можно найти ответы на все наши вопросы.

Если хочется получить точечную оценку, можно взять моду апостериорного распределения. Такой подход называют *байесом для бедных*. Обычно найти моду апостериорного распределения гораздо проще, чем сделать байесовский вывод.

$$\arg\max_\theta p(\theta \mid y, x) = \arg\max_\theta p(y \mid x, \theta)p(\theta) = \arg\max_\theta [\log p(y \mid x, \theta) + \log p(\theta)]$$

Получается, что при таком подходе мы максимизируем правдоподобие с некоторой добавкой. Эта добавка и представляет из себя регуляризацию.

Метод максимального правдоподобия — частный случай байесовских методов. В нём в качестве априорного распределения на параметры мы берём равномерное, а затем ищем моду апостериорного распределения.

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

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

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

**Задача 1.5.** *Покажите, что максимизация апостериорной вероятности $p(w \mid y, x)$ для модели линейной регрессии с нормальным априорным распределением эквивалентна решению задачи гребневой регрессии.*

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

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

Перейдем к логарифму и избавимся от константных членов:

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

В итоге получаем задачу гребневой регрессии

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

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

■

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

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

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

С точки зрения байесовского подхода [?] правильнее не искать моду[2] $w_*$ апостериорного распределения на параметрах и брать соответствующую ей модель $p(y \mid x, X, w_*)$, а устроить «взвешенное голосование» всех возможных моделей:

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

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

## §1.4 Наивный байесовский классификатор

Как было сказано ранее, при применении байесовского классификатора необходимо решить задачу восстановления плотности $p_y(x)$ для каждого класса $y \in \mathbb{Y}$. Данная задача является довольно трудоёмкой и не всегда может быть решена, особенно

[2] Мода — точка максимума плотности.

в случае большого количества признаков, — в частности, если объектами являются тексты, приходится работать с крайне большим числом признаков, и восстановление плотности многомерного распределения не представляется возможным.

Для разрешения этой проблемы сделаем предположение о независимости признаков. В этом случае функция правдоподобия класса $y$ для объекта $x = (x_1, \ldots, x_d)$ может быть представлена в следующем виде:

$$p(x \mid y) = \prod_{j=1}^{d} p(x_j \mid y),$$

где $p(x_j \mid y)$ — одномерная плотность распределения $j$-ого признака объектов класса $y \in Y$. В этом случае формула байесовского решающего правила примет следующий вид:

$$a(x) = \arg\max_{y \in Y} p(y \mid x) = \arg\max_{y \in \mathbb{Y}} \left( \ln p(y) + \sum_{j=1}^{d} \ln p(x_j \mid y) \right).$$

Предположение о независимости признаков существенно облегчает задачу, поскольку вместо решения задачи восстановления $d$-мерной плотности необходимо решить $d$ задач восстановления одномерных плотностей. Полученный классификатор называется *наивным байесовским классификатором*.

Плотности отдельных признаков могут быть восстановлены различными способами (параметрическими и непараметрическими). Среди параметрических способов чаще всего используются нормальное распределение (для вещественных признаков), распределение Бернулли и мультиномиальное распределение (для дискретных признаков), благодаря которым получаются различные применяющиеся на практике модели.