Семинары по композиционным методам

Евгений Соколов
sokolov.evg@gmail.com

30 марта 2014 г.

3 Композиционные методы машинного обучения

Ранее мы изучили различные вариации бустинга, которые жадно строили линейную комбинацию алгоритмов, повышая тем самым их качество. При этом каждый следующий алгоритм строился так, чтобы исправлять ошибки уже построенной композиции. Другим подходом к построению композиций является *бэггинг*, который независимо строит несколько алгоритмов и усредняет их ответы. Сначала мы рассмотрим инструмент, который поможет нам в анализе бэггинга — декомпозицию ошибки на компоненты смещения и разброса (bias-variance decomposition) — а затем перейдем к рассмотрению методов.

§3.1 Bias-Variance decomposition

3.1.1 Разложение в общем виде

Пусть задана выборка $X^\ell = \{x_1, \dots, x_\ell\} \subset \mathbb{X}$ с вещественными ответами $\{y_1, \dots, y_\ell\} \subset \mathbb{Y} = \mathbb{R}$ (рассматриваем задачу регрессии). Будем считать, что на пространстве всех объектов и ответов $\mathbb{X} \times \mathbb{Y}$ существует распределение $p(x, y)$, из которого и сгенерирована выборка X^ℓ и ответы на ней.

Рассмотрим квадратичную функцию потерь

$$L(y, a) = \big(y - a(x)\big)^2$$

и соответствующий ей среднеквадратичный риск

$$R(a) = \mathbb{E}_{x,y}\Big[\big(y - a(x)\big)^2\Big] = \int_{\mathbb{X}}\int_{\mathbb{Y}} p(x, y)\big(y - a(x)\big)^2 dxdy.$$

Ранее[1] мы показали, что минимум среднеквадратичного риска достигается на функции, возвращающей условное матожидание ответа при фиксированном объекте:

$$a^*(x) = \mathbb{E}[y \,|\, x] = \int_{\mathbb{Y}} yp(y \,|\, x)dx = \arg\min_a R(a).$$

[1] См. семинары по байесовским методам.

Для того, чтобы построить идеальную функцию регрессии, необходимо знать распределение на объектах и ответах $p(x, y)$, что, как правило, невозможно. На практике выбирается некоторый *метод обучения* $\mu : 2^{\mathbb{X}} \to \mathcal{A}$[2], который произвольной конечной обучающей выборке ставит в соответствие некоторый алгоритм из семейства $\mathcal{A}$. В качестве меры качества метода обучения можно взять усредненный по всем выборкам: среднеквадратичный риск алгоритма, выбранного методом по выборке:

$$L(\mu) = \mathbb{E}_{X^\ell}\Big[\mathbb{E}_{x,y}\Big[\big(y - \mu(X^\ell)\big)^2\Big]\Big]. \tag{3.1}$$

Здесь матожидание $\mathbb{E}_{X^\ell}[\cdot]$ берется по всем возможным выборкам $\{x_1, \ldots, x_\ell\}$ из распределения $\prod_{i=1}^{\ell} p(x_i, y_i)$.

Ранее[3] мы показали, что среднеквадратичный риск на фиксированной выборке X^ℓ можно расписать как

$$\mathbb{E}_{x,y}\Big[\big(y - \mu(X^\ell)\big)^2\Big] = \mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \mid x]\big)^2\Big] + \mathbb{E}_{x,y}\Big[\big(\mathbb{E}[y \mid x] - \mu(X^\ell)\big)^2\Big].$$

Подставим это представление в (3.1):

$$\begin{aligned} L(\mu) &= \mathbb{E}_{X^\ell}\Big[\underbrace{\mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \mid x]\big)^2\Big]}_{\text{не зависит от } X^\ell} + \mathbb{E}_{x,y}\Big[\big(\mathbb{E}[y \mid x] - \mu(X^\ell)\big)^2\Big]\Big] = \\ &= \mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \mid x]\big)^2\Big] + \mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\big(\mathbb{E}[y \mid x] - \mu(X^\ell)\big)^2\Big]\Big]. \end{aligned} \tag{3.2}$$

Преобразуем второе слагаемое:

$$\begin{aligned} &\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\big(\mathbb{E}[y \mid x] - \mu(X^\ell)\big)^2\Big]\Big] = \\ &= \mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\big(\mathbb{E}[y \mid x] - \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big] + \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big] - \mu(X^\ell)\big)^2\Big]\Big] = \\ &= \mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\underbrace{\big(\mathbb{E}[y \mid x] - \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big]\big)^2}_{\text{не зависит от } X^\ell}\Big]\Big] + \mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\big(\mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big] - \mu(X^\ell)\big)^2\Big]\Big] + \\ &\quad + 2\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\big(\mathbb{E}[y \mid x] - \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big]\big)\big(\mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big] - \mu(X^\ell)\big)\Big]\Big]. \end{aligned} \tag{3.3}$$

Покажем, что последнее слагаемое обращается в нуль:

$$\begin{aligned} &\mathbb{E}_{X^\ell}\Big[\big(\mathbb{E}[y \mid x] - \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big]\big)\big(\mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big] - \mu(X^\ell)\big)\Big] = \\ &\quad = \big(\mathbb{E}[y \mid x] - \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big]\big)\mathbb{E}_{X^\ell}\Big[\mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big] - \mu(X^\ell)\Big] = \\ &\quad = \big(\mathbb{E}[y \mid x] - \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big]\big)\Big[\mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big] - \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big]\Big] = \\ &\quad = 0. \end{aligned}$$

Учитывая это, подставим (3.3) в (3.2):

$$\begin{aligned} L(\mu) = &\underbrace{\mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \mid x]\big)^2\Big]}_{\text{шум}} + \\ &+ \underbrace{\mathbb{E}_{x,y}\Big[\big(\mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big] - \mathbb{E}[y \mid x]\big)^2\Big]}_{\text{смещение}} + \underbrace{\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\big(\mu(X^\ell) - \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big]\big)^2\Big]\Big]}_{\text{дисперсия}}. \end{aligned} \tag{3.4}$$

[2] Через $2^{\mathbb{X}}$ обозначается множество всех подмножеств множества $\mathbb{X}$.

[3] Опять же, см. семинары по байесовским методам.

Рассмотрим подробнее компоненты полученного разложения ошибки. Первая компонента характеризует *шум* в данных и равна ошибке идеального алгоритма. Невозможно построить алгоритм, имеющий меньшую среднеквадратичную ошибку. Вторая компонента характеризует *смещение (bias)* метода обучения, то есть отклонение среднего ответа обученного алгоритма от ответа идеального алгоритма. Третья компонента характеризует *дисперсию (variance)*, то есть разброс ответов обученных алгоритмов относительно среднего ответа.

Смещение показывает, насколько хорошо с помощью данных метода обучения и семейства алгоритмов можно приблизить оптимальный алгоритм. Как правило, смещение маленькое у сложных семейств (например, у деревьев) и большое у простых семейств (например, линейных классификаторов). Дисперсия показывает, насколько сильно может изменяться ответ обученного алгоритма в зависимости от выборки — иными словами, она характеризует чувствительность метода обучения к изменения в выборке. Как правило, простые семейства имеют маленькую дисперсию, а сложные семейства — большую дисперсию.

3.1.2 Разложение для метода k ближайших соседей

В качестве примера рассмотрим разложение ошибки для метода k ближайших соседей [1]. Будем считать, что выборка X^ℓ фиксирована и не является случайной, а случайность присутствует лишь в ответах:

$$p(y_i \,|\, x_i) = \mathcal{N}(f(x_i), \sigma^2),$$

где $f(x)$ — истинная зависимость. Вспомним, что метод k ближайших соседей находит к данному объекту k ближайших его соседей из обучающей выборки, и возвращает средний ответ на этих объектах:

$$a(x; X^\ell) = \frac{1}{k}\sum_{i=1}^{k} y_{(i)},$$

где $y_{(i)}$ — ответ на i-м по счету ближайшем к x объекте из X^ℓ.

Задача 3.1. *Найдите шумовую компоненту из разложения (3.4) для метода k ближайших соседей.*

Решение. Найдем сначала оптимальный алгоритм:

$$\mathbb{E}[y \,|\, x] = \mathbb{E}[\mathcal{N}(f(x), \sigma^2)] = f(x).$$

Вычислим теперь шумовую компоненту:

$$\mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y \,|\, x]\big)^2\Big] = \mathbb{E}_{x,y}\Big[\big(y - f(x)\big)^2\Big] =$$
$$= \mathbb{E}_{x,y}\Big[\big(f(x) + \mathcal{N}(0, \sigma^2) - f(x)\big)^2\Big] = \mathbb{E}_{x,y}\Big[\big(\mathcal{N}(0, \sigma^2)\big)^2\Big] = \sigma^2.$$

■

Задача 3.2. *Найдите смещение из разложения (3.4) для метода k ближайших соседей.*

Решение.

$$\mathbb{E}_{x,y}\Big[\big(\mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big] - \mathbb{E}[y \,|\, x]\big)^2\Big] = \mathbb{E}_{x,y}\Big[\big(\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[y_{(i)}] - f(x)\big)^2\Big] =$$

$$= \mathbb{E}_{x,y}\Big[\big(\frac{1}{k}\sum_{i=1}^{k} f(x_i) - f(x)\big)^2\Big].$$

Мы получили, что смещение равно матожиданию квадратичной ошибки метода k ближайших соседей, обученного по нашей выборке X^ℓ. Обратим внимание, что здесь измеряется ошибка идеального метода ближайших соседей, который использует незашумленные объекты.

■

Задача 3.3. *Найдите дисперсию из разложения* (3.4) *для метода k ближайших соседей.*

Решение.

$$\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\Big(\mu(X^\ell) - \mathbb{E}_{X^\ell}\big[\mu(X^\ell)\big]\Big)^2\Big]\Big] = \mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k} y_{(i)} - \mathbb{E}_{X^\ell}\Big[\frac{1}{k}\sum_{i=1}^{k} y_{(i)}\Big]\Big)^2\Big]\Big] =$$

$$= \mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k} y_{(i)} - \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[y_{(i)}]\Big)^2\Big]\Big] =$$

$$= \mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k} y_{(i)} - \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)})\Big)^2\Big]\Big] =$$

$$= \mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k} f(x_{(i)}) + \frac{1}{k}\sum_{i=1}^{k}\mathcal{N}(0,\sigma^2) - \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)})\Big)^2\Big]\Big] =$$

$$= \mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k}\mathcal{N}(0,\sigma^2)\Big)^2\Big]\Big] =$$

$$= \mathbb{E}_{x,y}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k}\mathcal{N}(0,\sigma^2)\Big)^2\Big] =$$

$$= \frac{\sigma^2}{k}.$$

Мы получили, что дисперсия метода k ближайших соседей обратно пропорциональна параметру k.

■

Получаем разложение для ошибки метода k ближайших соседей:

$$L(\mu_{\text{KNN}}) = \sigma^2 + \mathbb{E}_{x,y}\Big[\big(\frac{1}{k}\sum_{i=1}^{k} f(x_i) - f(x)\big)^2\Big] + \frac{\sigma^2}{k}.$$

Первое слагаемое характеризует уровень шума в данных и никак не может быть устранено. Второе и третье же слагаемые зависят от параметра алгоритма — числа соседей k. Если k велико, то алгоритм будет выдавать практически один и тот

же ответ во всем пространстве, из-за чего смещение будет большим; в то же время дисперсия будет маленькой, поскольку она обратно пропорциональна k. И наоборот, если k мало, то метод сможет восстановить очень сложную функцию, и смещение будет небольшим, но при этом дисперсия будет близка к σ^2. На практике нужно выбрать параметр k так, чтобы сумма этих двух слагаемых была минимальна — иными словами, найти компромисс между смещением и дисперсией.

§3.2 Бэггинг

Пусть имеется некоторый метод обучения $\mu(X^\ell)$. Построим на его основе метод $\tilde{\mu}(X^\ell)$, который генерирует случайную подвыборку $\tilde{X}^\ell \subset X^\ell$ с помощью бутстрэппинга, и подает ее на вход метода μ: $\tilde{\mu}(X^\ell) = \mu(\tilde{X}^k)$. В *бэггинге (bagging, bootstrap aggregation)* предлагается обучить некоторое число алгоритмов $b_n(x)$ с помощью метода $\tilde{\mu}$, и построить итоговую композицию как среднее данных базовых алгоритмов:

$$a_N(x) = \frac{1}{N}\sum_{n=1}^{N} b_n(x) = \frac{1}{N}\sum_{n=1}^{N} \tilde{\mu}(X^\ell)(x).$$

Задача 3.4. *Найдите смещение из разложения* (3.4) *для бэггинга.*

Решение.

$$\begin{aligned}
&\mathbb{E}_{x,y}\Big[\Big(\mathbb{E}_{X^\ell}\Big[\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X^\ell)(x)\Big] - \mathbb{E}[y\,|\,x]\Big)^2\Big] = \\
&\quad = \mathbb{E}_{x,y}\Big[\Big(\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x) - \mathbb{E}[y\,|\,x]\big]\Big)^2\Big] = \\
&\quad = \mathbb{E}_{x,y}\Big[\big(\mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x) - \mathbb{E}[y\,|\,x]\big]\big)^2\Big] = \\
&\quad = \mathbb{E}_{x,y}\Big[\big(\mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big] - \mathbb{E}[y\,|\,x]\big)^2\Big].
\end{aligned}$$

Мы получили, что смещение композиции, полученной с помощью бэггинга, совпадает со смещением одного базового алгоритма. Таким образом, бэггинг не ухудшает смещенность модели.

■

Задача 3.5. *Найдите дисперсию из разложения* (3.4) *для бэггинга.*

Решение. Запишем выражение для дисперсии композиции, обученной с помощью бэггинга:

$$\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\Big(\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\Big[\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X^\ell)(x)\Big]\Big)^2\Big]\Big].$$

Рассмотрим выражение, стоящее под матожиданиями:

$$\Big(\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\Big[\frac{1}{N}\sum_{n=1}^{N}\tilde{\mu}(X^\ell)(x)\Big]\Big)^2 =$$
$$= \frac{1}{N^2}\Big(\sum_{n=1}^{N}\Big[\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big]\Big)^2 =$$
$$= \frac{1}{N^2}\sum_{n=1}^{N}\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)^2 +$$
$$+ \frac{1}{N^2}\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)$$

Возьмем теперь матожидания от этого выражения, учитывая, что все базовые алгоритмы одинаково распределены относительно X^ℓ:

$$\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\frac{1}{N^2}\sum_{n=1}^{N}\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)^2 +$$
$$+ \frac{1}{N^2}\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)\Big]\Big] =$$
$$= \frac{1}{N^2}\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\sum_{n=1}^{N}\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)^2\Big]\Big] +$$
$$+ \frac{1}{N^2}\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\sum_{n_1\neq n_2}\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)\times$$
$$\times\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)\Big]\Big] =$$
$$= \frac{1}{N}\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)^2\Big]\Big] +$$
$$+ \frac{N(N-1)}{N^2}\mathbb{E}_{x,y}\Big[\mathbb{E}_{X^\ell}\Big[\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)\times$$
$$\times\Big(\tilde{\mu}(X^\ell)(x) - \mathbb{E}_{X^\ell}\big[\tilde{\mu}(X^\ell)(x)\big]\Big)\Big]\Big]$$

Первое слагаемое — это дисперсия одного базового алгоритма, деленная на длину композиции N. Второе — ковариация между двумя базовыми алгоритмами. Мы видим, что если базовые алгоритмы некоррелированы, то дисперсия композиции в N раз меньше дисперсии отдельных алгоритмов. Если же корреляция имеет место, то уменьшение дисперсии может быть гораздо менее существенным.

■

§3.3 Случайные леса

Как мы выяснили, бэггинг позволяет объединить несмещенные, но чувствительные к обучающей выборке алгоритмы в несмещенную композицию с низкой дисперсией. Хорошим семейством базовых алгоритмов здесь являются решающие деревья —

они достаточно сложны и могут достигать нулевой ошибки на любой выборке (следовательно, имеют низкое смещение), но в то же время легко переобучаются.

Метод *случайных лесов* [2] основан на бэггинге над решающими деревьями. Выше мы отметили, что бэггинг сильнее уменьшает дисперсию базовых алгоритмов, если они слабо коррелированы. В случайных лесах корреляция между деревьями понижается путем рандомизации. Вспомним, что при построении дерева последовательно происходит разделение вершин до тех пор, пока не будет достигнуто идеальное качество на обучении. Каждая вершина разбивает выборку по одному из признаков относительно некоторого порога. В случайных лесах признак, по которому производится разбиение, выбирается не из всех возможных признаков, а лишь из их случайного подмножества размера k.

Рекомендуется в задачах классификации брать $k = \lfloor\sqrt{d}\rfloor$, а в задачах регрессии — $k = \lfloor d/3\rfloor$, где d — число признаков. Также рекомендуется в задачах классификации строить каждое дерево до тех пор, пока в каждом листе не окажется по одному объекту, а в задачах регрессии — пока в каждом листе не окажется по пять объектов.

Случайные леса — один из самых сильных методов построения композиций. На практике он может работать немного хуже градиентного бустинга, но при этом он гораздо более прост в реализации.

Список литературы

[1] *Hastie, T., Tibshirani, R., Friedman, J.* (2001). The Elements of Statistical Learning. // Springer, New York.

[2] *Breiman, Leo* (2001). Random Forests. // Machine Learning, 45(1), 5–32.