Семинар 2.
ММП, весна 2013
26 февраля

Илья Толстихин
iliya.tolstikhin@gmail.com

Темы семинара:

- Композиции и методы их построения;
- 'bias-variance tradeoff";
- Bagging.

1 Бэггинг

Начнем с рассмотрения задачи регрессии $\mathbb{Y} = \mathbb{R}$ с квадратичной функцией потерь:

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

Мы будем работать в рамках вероятностной постановки задачи обучения по прецедентам: то есть на декартовом произведении $\mathbb{X} \times \mathbb{Y}$ задана неизвестная нам вероятностная мера $P(X, Y)$. Далее мат. ожидание по этой мере будем обозначать с помощью $\mathbb{E}_{x,y}$, то есть для абсолютно непрерывного распределения $P(X, Y)$ с плотностью $p(X, Y)$ можно записать $\mathbb{E}_{x,y}[f(x, y)] = \int_{z \in \mathbb{X}\times\mathbb{Y}} f(z)p(z)dz$. Представим, что мы с помощью обучающей выборки $X^\ell = \{x_i, y_i\}_{i=1}^\ell$ выбрали некоторую функцию $h(x)$ и вычисляем *средний риск*, связанный с ее использованием:

$$\mathbb{E}_{x,y}\left[\big(y - h(x)\big)^2\right].$$

Мы знаем, что функционал среднего квадратичного риска достигает своего минимума на *функции регрессии* — условном математическом ожидании $\mathbb{E}[y|x]$. Вспомним доказательство.

$$\mathbb{E}_{x,y}\left[\big(y - h(x)\big)^2\right] = \mathbb{E}_{x,y}\left[\big(y - \mathbb{E}[y|x] + \mathbb{E}[y|x] - h(x)\big)^2\right] =$$
$$= \mathbb{E}_{x,y}\left[\big(y - \mathbb{E}[y|x]\big)^2\right] + \mathbb{E}_{x,y}\left[\big(\mathbb{E}[y|x] - h(x)\big)^2\right] + 2\mathbb{E}_{x,y}\left[\big(y - \mathbb{E}[y|x]\big)\big(\mathbb{E}[y|x] - h(x)\big)\right]. \tag{1}$$

Введем функцию $\mathbb{E}_{y|x}[f(x, y)]$ случайной величины x следующим образом:

$$\mathbb{E}_{y|x}[f(x, y)] = \int_{y\in\mathbb{Y}} f(x, y)p(y|x)dy.$$

(Заметим, что $\mathbb{E}[y|x] = \int yp(y|x)dy = \mathbb{E}_{y|x}[y]$). Тогда, воспользовавшись теоремой Фубини, мы получаем

$$\mathbb{E}_{x,y}[f(x,y)] = \int_{z\in\mathbb{X}\times\mathbb{Y}} f(z)p(z)dz = \int_{x\in\mathbb{X}}\left[\int_{y\in\mathbb{Y}} f(x,y)p(y|x)dy\right]p(x)dx = \mathbb{E}_x\big[\mathbb{E}_{y|x}[f(x,y)]\big].$$

Применим полученное тождество к последнему слагаемому неравенства (1) и получим

$$\mathbb{E}_{x,y}\left[\big(y-\mathbb{E}[y|x]\big)\big(\mathbb{E}[y|x]-h(x)\big)\right] = \mathbb{E}_x\mathbb{E}_{y|x}\left[\big(y-\mathbb{E}[y|x]\big)\big(\mathbb{E}[y|x]-h(x)\big)\right] = \\ = \mathbb{E}_x\left[\mathbb{E}_{y|x}\big[y-\mathbb{E}[y|x]\big]\big(\mathbb{E}[y|x]-h(x)\big)\right] = 0,$$

поскольку выражение $\mathbb{E}[y|x]-h(x)$ не зависит от y.

Таким образом

$$\mathbb{E}_{x,y}\left[\big(y-h(x)\big)^2\right] = \mathbb{E}_{x,y}\left[\big(y-\mathbb{E}[y|x]\big)^2\right] + \mathbb{E}_{x,y}\left[\big(\mathbb{E}[y|x]-h(x)\big)^2\right]. \tag{2}$$

Итак, выбрав в качестве $h(x)$ функцию регрессии $\mathbb{E}[y|x]$, мы получим минимальный достижимый риск $\mathbb{E}_{x,y}\left[\big(y-\mathbb{E}[y|x]\big)^2\right]$ — он обусловлен наличием шума в выборке.

Задача. *В каком случае минимальный достижимый средний риск в задаче восстановления регрессии с квадратичной функцией потерь равен нулю?*

В цепочке рассуждений, представленной выше, мы никак не выделяли факт зависимости выбранной нами функции $h(x)$ от обучающей выборки X^ℓ. В прикладных задачах нас все же будет интересовать поведение получаемой нами с помощью того или иного метода обучения $\mu\colon X^\ell \to \mathcal{H} = \{h\colon \mathbb{X}\to\mathbb{Y}\}$ функции $h^\ell(x)$ в среднем при варьировании обучающих выборок (здесь верхний индекс ℓ указывает на зависимость функции h^ℓ от обучающей выборки). Для этого рассмотрим следующую характеристику используемого метода обучения μ:

$$\mathcal{L}(\mu) = \mathbb{E}_{X^\ell}\left[\mathbb{E}_{x,y}\left[\big(y-h^\ell(x)\big)^2\right]\right],\ \text{где } h^\ell = \mu\left(X^\ell\right),$$

отражающую среднее (относительно случайных обучающих выборок) значение риска функции, получаемой при использовании метода обучения μ. Здесь $\mathbb{E}_{X^\ell}[f(X^\ell)]$ обозначает математическое ожидание функции f относительно вероятного распределения на всевозможных обучающих выборках $X^\ell \in (\mathbb{X}\times\mathbb{Y})^\ell$ длинны ℓ.

Воспользовавшись (2), мы получаем

$$\mathbb{E}_{X^\ell}\left[\mathbb{E}_{x,y}\left[\big(y-h^\ell(x)\big)^2\right]\right] = \mathbb{E}_{x,y}\left[\big(y-\mathbb{E}[y|x]\big)^2\right] + \mathbb{E}_{X^\ell}\left[\mathbb{E}_{x,y}\left[\big(\mathbb{E}[y|x]-h^\ell(x)\big)^2\right]\right].$$

Попробуем изучить поведение функции $h^\ell(x)$ относительно величины $\mathbb{E}_{X^\ell}\left[h^\ell(x)\right]$, преобразовав вид второго слагаемого:

$$\mathbb{E}_{X^\ell}\left[\mathbb{E}_{x,y}\left[\big(\mathbb{E}[y|x]-h^\ell(x)\big)^2\right]\right] = \mathbb{E}_{x,y}\left[\mathbb{E}_{X^\ell}\left[\big(\mathbb{E}[y|x]-\mathbb{E}_{X^\ell}\left[h^\ell(x)\right]+\mathbb{E}_{X^\ell}\left[h^\ell(x)\right]-h^\ell(x)\big)^2\right]\right]. \tag{3}$$

Снова воспользовавшись теоремой Фубини:

$$\mathbb{E}_{X^\ell}\Big[\big(\mathbb{E}[y|x]-\mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\big)\left(\mathbb{E}_{X^\ell}\left[h^\ell(x)\right]-h^\ell(x)\right)\Big] = \\ = \big(\mathbb{E}[y|x]-\mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\big)\,\mathbb{E}_{X^\ell}\Big[\left(\mathbb{E}_{X^\ell}\left[h^\ell(x)\right]-h^\ell(x)\right)\Big] = 0,$$

мы получим:

$$\mathbb{E}_{X^\ell}\left[\left(\mathbb{E}[y|x] - h^\ell(x)\right)^2\right] =$$
$$= \underbrace{\left(\mathbb{E}[y|x] - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\right)^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{X^\ell}\left[\left(h^\ell(x) - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\right)^2\right]}_{\text{variance}}.$$

Итак, мы видим, что среднее (относительно случайного выпадения обучающей выборки) значение квадратичной функции потерь представляется в виде суммы двух неотрицательных слагаемых. Первое слагаемое — отклонение — характеризует отклонение усредненного по всевозможным обучающим выборкам ответа получаемых нами с помощью метода обучения μ функций h^ℓ от ответа функции регрессии $\mathbb{E}[y|x]$, которую мы стремимся приблизить. Второе слагаемое — дисперсия — характеризует разброс ответов функций h^ℓ вокруг их усредненного ответа $\mathbb{E}_{X^\ell}\left[h^\ell(x)\right]$. Эту величину можно интерпретировать как чувствительность используемой нами модели к изменениям обучающей выборки.

Подставив полученный результат в (3), мы получаем следующее разложение среднего риска, связанного с использованием метода обучения μ:

$$\mathcal{L}(\mu) =$$
$$= \underbrace{\mathbb{E}_{x,y}\left[(y - \mathbb{E}[y|x])^2\right]}_{\text{noise}} + \underbrace{\mathbb{E}_x\left[\left(\mathbb{E}[y|x] - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\right)^2\right]}_{(\text{bias})^2} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_{X^\ell}\left[\left(h^\ell(x) - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\right)^2\right]\right]}_{\text{variance}}. \tag{4}$$

Обычно у больших и гибких моделей низкое значение отклонения, но высокая дисперсия. У простых моделей наоборот — отклонение высоко, но дисперсия мала. Для достижения минимума риска приходится искать копромисс между этими двумя величинами. Проиллюстрируем эти рассуждения на следующем примере.

Bias-Variance разложение для KNN. Предположим, что в задаче регрессии $\mathbb{Y} = \mathbb{R}$ распределение $P(X, Y)$ таково, что $Y = f(X) + \varepsilon$, где случайная величина ε независима от X, $\mathbb{E}[\varepsilon] = 0$ и $\text{Var}[\varepsilon] = \sigma^2$. Также для простоты предположим, что объекты $x_1, \ldots, x_\ell$ обучающей выборки зафиксированы и случайность обучающей выборки обусловлена лишь случайностью ответов на точках обучающей выборки. Для решения этой задачи мы будем использовать алгоритм K-ближайших соседей, то есть

$$h(x) = \frac{1}{K}\sum_{k=1}^{K} y_{(k)},$$

где $y_{(k)}$ — ответ на k-ом соседе x из обучающей выборки.

Задача. *Вычислите разложение (4) для квадратичной функции потерь при использовании алгоритма K-ближайших соседей.*

Бэггинг. Процедура *бэггинга* заключается в следующем. Мы будем генерировать M новых обучающих выборок $\tilde{X}_m^\ell$, $m = 1, \ldots, M$, той же длины ℓ, вытягивая их из равномерного распределения на X^ℓ (будем на каждом шаге тянуть один объект

с ответом на нем из обучающей выборки *с возвращением*). На каждой полученной обучающей выборке $\tilde{X}_m^\ell$ мы будем обучать функцию $h_m = \mu\left(\tilde{X}_m^\ell\right)$. Затем из полученных таким образом функций построим итоговую функцию

$$h(x) = \frac{1}{M}\sum_{m=1}^{M} h_m(x).$$

Термин bagging происходит от словосочетания bootstrap aggregation.

Задача. *Найдите, чему равны шум, отклонение и дисперсия в разложении* (4) *для композиции, полученной с помощью бэггинга.*

Решение. Рассмотрим выражение, стоящее под квадратом в слагаемом, соответствующем отклонению:

$$\mathbb{E}[y|x] - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right] = \mathbb{E}[y|x] - \mathbb{E}_{X^\ell}\left[\frac{1}{M}\sum_{m=1}^{M} h_m(x)\right] = \frac{1}{M}\sum_{m=1}^{M}\Big(\mathbb{E}[y|x] - \mathbb{E}_{X^\ell}\left[h_m(x)\right]\Big).$$

Мы представили это выражение в виде среднего значения отклонений. При этом функция h_m зависит от выборки $\tilde{X}_m^\ell$, которую мы вытянули из первоначальной. Поскольку ответы функций $h_m(x)$, $m = 1, \ldots, m$, в точке x распределены одинаково (относительно случайной реализации обучающей выборки), мы заключаем, что

$$\mathbb{E}_{X^\ell}\left[h_m(x)\right] = \mathbb{E}_{X^\ell}\left[h_{m'}(x)\right],$$

то есть для произвольного m

$$\mathbb{E}[y|x] - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right] = \mathbb{E}[y|x] - \mathbb{E}_{X^\ell}\left[h_m(x)\right].$$

Таким образом, отклонение композиции, получаемой с помощью бэггинга, не отличается от отклонения любой из функций, вошедших в композицию.

Теперь перейдем к рассмотрению дисперсии усредненной функции. Рассмотрим выражение под квадратом в последнем слагаемом:

$$\mathbb{E}_{X^\ell}\left[\left(h^\ell(x) - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\right)^2\right] = \mathbb{E}_{X^\ell}\left[\left(\frac{1}{M}\sum_{m=1}^{M} h_m(x) - \mathbb{E}_{X^\ell}\frac{1}{M}\sum_{m=1}^{M} h_m(x)\right)^2\right] =$$

$$= \frac{1}{M^2}\mathbb{E}_{X^\ell}\left[\left(\sum_{m=1}^{M}\Big(h_m(x) - \mathbb{E}_{X^\ell}\left[h_m(x)\right]\Big)\right)^2\right] =$$

$$= \frac{1}{M^2}\mathbb{E}_{X^\ell}\left[\sum_{m=1}^{M}\Big(h_m(x) - \mathbb{E}_{X^\ell}\left[h_m(x)\right]\Big)^2 + \sum_{i\neq j}\Big(h_i(x) - \mathbb{E}_{X^\ell}\left[h_i(x)\right]\Big)\Big(h_j(x) - \mathbb{E}_{X^\ell}\left[h_j(x)\right]\Big)\right] =$$

$$= \frac{1}{M^2}\sum_{m=1}^{M}\mathrm{Var}_{X^\ell}[h_m(x)] + \frac{1}{M^2}\sum_{i\neq j}\mathrm{Cov}_{X^\ell}[h_i(x), h_j(x)].$$

Вновь учитывая одинаковое распределение ответов $h_m(x)$, мы приходим к следующему виду дисперсии композиции, построенной бэггингом:

$$\mathbb{E}_{X^\ell}\left[\left(h^\ell(x) - \mathbb{E}_{X^\ell}\left[h^\ell(x)\right]\right)^2\right] = \frac{1}{M}\mathrm{Var}_{X^\ell}[h_m(x)] + \frac{1}{M^2}\sum_{i\neq j}\mathrm{Cov}_{X^\ell}[h_i(x), h_j(x)].$$

Итак, мы видим, что если базовые функции h_m, получаемые бэггингом на очередных выборках $\tilde{X}_m^\ell$, некоррелированы, то бэггинг уменьшает дисперсию в M раз.

Стоит отметить, что в действительности бэггинг уменьшает дисперсию не так сильно, поскольку базовые функции как правило коррелируют. Также отклонение композиции бэггинга немного подрастает по сравнению с методом μ, примененным ко всей обучающей выборке. Это объясняется тем, что фактически бэггинг уменьшает эффективный размер обучающей выборки, на которой запускается процедура обучения μ, что связано с семплированием из равномерного распределения на обучающей выборке: в каждую из выборок $\tilde{X}_m^\ell$ будут входить повторяющиеся объекты.

Описанный эффект также объясняет, почему бэггинг, будучи примененным к простыми методами обучения, мало чувствительным к небольшим изменениям обучающей выборки (таким как decision stump), может существенно ухудшить качество на контрольной выборке. Уменьшение дисперсии композиции в этом случае не будет существенной по сравнению с итак малым значением дисперсии отдельной функции, настроенной по всей обучающей выборке. Однако, отклонение композиции может существенно увеличиться по сравнению с отклонением отдельной функции, обученной по всей обучающей выборке.

Замечание. В отличие от бэггинга, который наращивает базовые функции независимо друг от друга, бустинг наращивает их последовательно и процедура построения новой базовой функции зависит от прошлых шагов. С одной стороны, такая адаптивная стратегия не позволяет эффективно распараллелить бустинг на этапе обучения, запуская обучение каждой базовой функции в отдельном потоке. С другой стороны, адаптивная стратегия бустинга позволяет уменьшить отклонение композиции. По этой причине бустинг часто работает лучше бэггинга в прикладных задачах. Метод Random Forest [1] сочетает бэггинг решающих деревьев с дополнительной рандомизацией, семплируя новые признаковые подпространства при построении каждой внутренней вершины дерева, что ведет к уменьшению корреляции между получаемыми таким образом базовыми функциями. На многих прикладных задачах Random Forest показывает результаты, сравнимые с бустингом, и в то же время его гораздо проще реализовать.

Список литературы

[1] *Hastie, T., R. Tibshirani, and J. H. Friedman.* The Elements of Statistical Learning: Data Mining, Inference and Prediction. — Springer, 2001.

[2] *Bishop, C.* Pattern recognition and machine learning. — Springer, 2006.