Машинное обучение, ФКН ВШЭ
Семинар №8

1 Разложение на смещение и разброс

На лекции была выведена следующая формула, показывающая, как можно представить ошибку алгоритма регрессии в виде суммы трех компонент:

$$L(\mu) = \mathbb{E}_{x,y}\left[\mathbb{E}_X\left[(y - \mu(X)(x))^2\right]\right] =$$

$$\underbrace{\mathbb{E}_{x,y}\left[(y - \mathbb{E}\left[y|x\right])^2\right]}_{\text{шум}} + \underbrace{\mathbb{E}_x\left[(\mathbb{E}_X\left[\mu(X)(x)\right] - \mathbb{E}\left[y|x\right])^2\right]}_{\text{смещение}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_X\left[(\mu(X)(x) - \mathbb{E}_X\left[\mu(X)(x)\right])^2\right]\right]}_{\text{разброс}},$$

- $\mu(X)$ — алгоритм, обученный по выборке $X = \{(x_1, y_1), \dots (x_\ell, y_\ell)\}$;
- $\mu(X)(x)$ — ответ алгоритма, обученного по выборке X, на объекте x;
- $\mathbb{E}_X$ — мат. ожидание по всем возможным выборкам;
- $\mathbb{E}_X\left[\mu(X)(x)\right]$ — «средний» ответ алгоритма, обученного по всем возможным выборкам X, на объекте x.

С помощью этой формулы мы можем анализировать свойства алгоритма обучения модели μ, если зададим вероятностную модель порождения пар $p(x, y)$.

§1.1 Связь регуляризации с BVD

Чтобы лучше понять смысл трех компонент, входящих в разложение, рассмотрим одномерную линейную регрессию с L_2 регуляризатором.

Алгоритм обучения μ. В одномерной линейной регрессии зависимость целевого признака y от объекта x моделируется с помощью примитивной линейной функции $y = wx$. Оптимальный параметр w находится по выборке $X = \{(x_1, y_1), \dots (x_\ell, y_\ell)\}$ минимизацией

$$L(w) = \sum_{i=1}^{\ell}(y_i - wx_i)^2 + \lambda \cdot w^2.$$

При решении данной задачи оптимизации получается следующий алгоритм $\mu(X)$:

$$\mu(X)(x) = w(X)\, x, \quad w(X) = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \lambda}.$$

Обратите внимание, что мы здесь оптимизируем сумму ошибок. Если бы мы оптимизировали

$$L(w) = \frac{1}{\ell}\sum_{i=1}^{\ell}(y_i - wx_i)^2 + \lambda \cdot w^2,$$

тогда мы бы получили

$$w(X) = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \ell \cdot \lambda}.$$

Получается, что если мы не усредняем функцию потерь, тогда гиперпараметр λ будет функцией от ℓ.

Чтобы сделать разложение на смещение и разброс, нужно описать вероятностную модель данных. Это можно делать по-разному. Мы попробуем два набора из предпосылок.

Вероятностная модель №1. Выборка X составляется из ℓ независимых пар (x_i, y_i). Будем считать, что объекты x детерминированы, то есть это не случайные величины. Такая предпосылка упростит нам вычисления. Во второй модели мы от неё откажемся.

Бывают ли объекты x детерминированы в реальной жизни? Иногда бывают. Например, если мы анализируем цены на квартиры, мы можем собрать выборку из 100 квартир площадью 30 м^2, 100 площадью 35 м^2 и 100 площадью 40 м^2. Если мы соберём другую выборку из трёхсот квартир с такими же площадями, то значения x останутся прежними, а значения y поменяются.

Правильный ответ на объекте x определяется зашумлённой функцией $f(x)$: $y = f(x) + \varepsilon$. Будем предполагать, что шум пришёл к нам из какого-то распределения с нулевым средним и дисперсией σ^2. Запишем это как $\varepsilon \sim (0, \sigma^2)$. Иными словами, $y \sim (f(x), \sigma^2)$.

Задача 1.1. *Пусть истинная модель порождения данных выглядит как $f(x) = ax$. Найдите шумовую компоненту для одномерной линейной регрессии с L_2 регуляризатором.*

Решение. Вероятностная модель позволяет нам довольно легко найти математическое ожидание шумовой компоненты

$$\mathbb{E}\left[y|x\right] = f(x).$$

Тогда

$$\mathbb{E}_{x,y}\left[(y - \mathbb{E}\left[y|x\right])^2\right] = \mathbb{E}_{\varepsilon}\left[(f(x) + \varepsilon - f(x))^2\right] = \mathbb{E}_{\varepsilon}\left[\varepsilon^2\right] = \mathrm{Var}\left[\varepsilon\right] + \left(\mathbb{E}\varepsilon\right)^2 = \sigma^2 + 0 = \sigma^2.$$

■

Задача 1.2. *Найдите смещение алгоритма одномерной линейной регрессии с L_2 регуляризатором для $f(x) = ax$.*

Решение. Для начала найдем «средний» по всем выборкам ответ алгоритма на объекте x:

$$\mathbb{E}_X\left[\mu(X)(x)\right] = \mathbb{E}_X\left[w(X)\right]\, x.$$

Помним, что все x фиксированы и их смело можно выносить за математическое ожидание. Случайность в вычислениях есть только из-за ошибки ε

$$\mathbb{E}_X\left[w(X)\right] = \mathbb{E}_\varepsilon\left[\frac{\sum_i x_i(f(x_i)+\varepsilon_i)}{\sum_i x_i^2+\lambda}\right] = \mathbb{E}_\varepsilon\left[\frac{\sum_i x_i(ax_i+\varepsilon_i)}{\sum_i x_i^2+\lambda}\right] =$$
$$= a\cdot\frac{\sum_i x_i^2}{\sum_i x_i^2+\lambda} + \frac{\sum_i x_i\cdot\mathbb{E}\left[\varepsilon_i\right]}{\sum_i x_i^2+\lambda} = a\cdot\frac{\sum_i x_i^2}{\sum_i x_i^2+\lambda}.$$

Найдем смещение

$$\mathbb{E}_x\left[\mathbb{E}_X\left[\mu(X)(x)\right] - \mathbb{E}\left[y|x\right]\right] = a\cdot\frac{\sum_i x_i^2}{\sum_i x_i^2+\lambda}\cdot x - a\,x = a\cdot\frac{s}{s+\lambda}\cdot x - a\,x$$

Мы сняли математическое ожидание $\mathbb{E}_x$, так как считаем объекты детерминированными. За s мы обозначили величину $\sum_i x_i^2$.

Если $\dfrac{s}{s+\lambda} = 1$, тогда $\lambda = 0$ и смещение модели равно нулю. В этом случае мы оптимизируем MSE без регуляризатора и имеем дело с обычной линейной регрессией. Если $\lambda > 0$, тогда в модели появляется смещение. При этом смещёнными оказываются не только прогнозы модели, но и коэффициенты в ней. Найдём квадрат смещения, именно его мы будем использовать в разложении

$$\mathbb{E}_x\left[\left(\mathbb{E}_X\left[\mu(X)(x)\right] - \mathbb{E}\left[y|x\right]\right)^2\right] = \left(\frac{s}{s+\lambda}-1\right)^2 a^2\,x^2$$

■

Задача 1.3. *Найдите разброс алгоритма одномерной линейной регрессии с L_2 регуляризатором для $f(x) = ax$.*

Решение. Нам надо найти

$$\mathbb{E}_x\left[\mathrm{Var}_X\left[\mu(X)(x)\right]\right] = \mathbb{E}_x\left[\mathbb{E}_X\left[\left(\mu(X)(x) - \mathbb{E}_X\left[\mu(X)(x)\right]\right)^2\right]\right].$$

Найдём дисперсию алгоритма по обучающей выборке

$$\mathrm{Var}_X(\mu(X)(x)) = \mathrm{Var}_X\left[\frac{\sum_i x_i y_i}{\sum_i x_i^2+\lambda}x\right] = \mathrm{Var}_X\left[\frac{\sum_i x_i(f(x_i)+\varepsilon_i)}{\sum_i x_i^2+\lambda}x\right] =$$
$$= \mathrm{Var}_X\left[\frac{\sum_i a\,x_i^2}{\sum_i x_i^2+\lambda}\,x + \frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2+\lambda}x\right] = \mathrm{Var}_X\left[\frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2+\lambda}x\right] =$$
$$= \frac{1}{(\sum_i x_i^2+\lambda)^2}\,x^2\sum_i x_i^2\,\mathrm{Var}_\varepsilon(\varepsilon_i) = \frac{\sum_i x_i^2}{(\sum_i x_i^2+\lambda)^2}\,x^2\,\sigma^2$$

Получается, что

$$\mathbb{E}_x\left[\mathrm{Var}_X\left[\mu(X)(x)\right]\right] = \mathbb{E}_x\left[\mathbb{E}_X\left[\left(\mu(X)(x) - \mathbb{E}_X\left[\mu(X)(x)\right]\right)^2\right]\right] =$$
$$= \mathbb{E}_x\left[\frac{\sum_i x_i^2}{(\sum_i x_i^2+\lambda)^2}\,x^2\,\sigma^2\right] = \frac{\sum_i x_i^2}{(\sum_i x_i^2+\lambda)^2}\,x^2\,\sigma^2. = \frac{s}{(s+\lambda)^2}\,x^2\,\sigma^2.$$

Последнее математическое ожидание мы сняли, так как предполагаем, что x детерминированы.

■

Запишем получившееся разложение

$$\mathbb{E}_{x,y}\left[\mathbb{E}_X\left[(y-\mu(X)(x))^2\right]\right] = \sigma^2 + \left(\frac{s}{s+\lambda} - 1\right)^2 a^2\, x^2 + \frac{s}{(s+\lambda)^2}\, x^2\, \sigma^2.$$

С помощью гиперпараметра λ мы можем настраивать силу регуляризации. Чем больше это значение, тем меньше разброс модели и больше смещение.

Задача 1.4. *Подберите λ, которое будет давать нам оптимальный баланс между смещением и разбросом. Это значение должно улучшать качество прогноза с точки зрения среднеквадратичной ошибки. Правда ли, что λ отличается от нуля? Почему силу регуляризации нельзя подобрать по выборке, воспользовавшись получившейся формулой?*

Решение. Решим задачу оптимизации

$$\sigma^2 + \left(\frac{s}{s+\lambda} - 1\right)^2 a^2\, x^2 + \frac{s}{(s+\lambda)^2}\, x^2\, \sigma^2 \to \min_{\lambda}.$$

Сделаем замену переменной $\gamma = \dfrac{1}{s+\lambda}$

$$\sigma^2 + (s\,\gamma - 1)^2\; a^2\, x^2 + s\,\gamma^2\, x^2\, \sigma^2 \to \min_{\lambda}.$$

Перед нами парабола с ветвями вверх

$$(sa^2 + \sigma^2)\, x^2\, s\,\gamma^2 - 2\, s\, a^2\, x^2\, \gamma + a^2\, x^2 + \sigma^2$$

значит её минимум можно найти как

$$\gamma^* = \frac{2\, s\, a^2\, x^2}{2\,(s\, a^2 + \sigma^2)\, x^2\, s} = \frac{a^2}{s\, a^2 + \sigma^2} = \frac{1}{s + \sigma^2/a^2}.$$

Получается

$$\lambda^* = \frac{\sigma^2}{a^2}.$$

Получившаяся величина отличается от нуля, так как $\sigma^2 > 0$. Выходит, что модель, дающая лучший прогноз, оказывается смещённой. Гиперпараметр λ мы не можем подобрать по обучающей выборке, так как не знаем значения коэффициента a. Нам нужно оценить его с помощью нашей модели, оптимизация которой, в свою очередь, зависит от выбранного значения λ. Из-за этого приходится подбирать силу регуляризатора с помощью перебора.

■

Если в предпосылках у нас будут стохастические признаки x, вычисления станут более сложными, но зато результаты станут более общими.

§1.2 BVD для случайных признаков

Попробуем проделать всё то же самое для случайных признаков. Для простоты будем работать с линейной моделью без регуляризации. Для неё

$$\mu(X)(x) = w(X)\,x, \quad w(X) = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$

Вероятностная модель №2. Выборка X составляется из ℓ независимых пар (x_i, y_i). Будем считать, что объекты x генерируются из нормального распределения $x \sim p(x) = \mathcal{N}(0, \sigma_1^2)$. Правильный ответ на объекте x определяется зашумленной функцией $f(x) : y = f(x) + \varepsilon$, $\varepsilon \sim p(\varepsilon) = \mathcal{N}(0, \sigma_2^2)$. Иными словами, $y \sim p(y|x) = \mathcal{N}(f(x), \sigma_2^2)$.

Мы будем рассматривать два простых частных случая: $f(x) = ax$, когда модель зависимости отвечает искомой зависимости, и $f(x)$ — четная функция, т. е. $f(-x) = f(x)$.

Задача 1.5. *Найдите шумовую компоненту для одномерной линейной регрессии.*

Решение. Вычисления почти никак не изменятся. Так как распределение $p(y|x)$ нормальное, для него легко вычислить математическое ожидание:

$$\mathbb{E}\,[y|x] = f(x).$$

Тогда

$$\mathbb{E}_{x,y}\left[(y - \mathbb{E}\,[y|x])^2\right] = \mathbb{E}_{x,\varepsilon}\left[(f(x) + \varepsilon - f(x))^2\right] = \mathbb{E}_\varepsilon(\varepsilon^2) = \mathrm{Var}_\varepsilon(\varepsilon) + (\mathbb{E}\varepsilon)^2 = \sigma_2^2 + 0 = \sigma_2^2.$$

■

Задача 1.6. *Найдите смещение алгоритма одномерной линейной регрессии для $f(x) = ax$ и для произвольной четной $f(x)$.*

Решение. Для начала найдем «средний» по всем выборкам ответ алгоритма на объекте x:

$$\mathbb{E}_X\,[\mu(X)(x)] = \mathbb{E}_X\,[w(X)]\;x.$$

Итак, нам нужно найти «среднее» по всем выборкам значение коэффициента w:

$$\mathbb{E}_X\,[w(X)] = \int \frac{\sum_i x_i(f(x_i) + \varepsilon_i)}{\sum_i x_i^2} \prod_i \left(p(x_i)p(\varepsilon_i)\right)\, dx_1 \ldots dx_\ell\, d\varepsilon_1 \ldots d\varepsilon_\ell.$$

Здесь записан несобственный интеграл, в котором каждая переменная принимает значения от $-\infty$ до ∞. Значение этого интеграла определяется функцией $f(x)$ и не всегда вычисляется аналитически. В случае, когда истинная зависимость в данных линейная, мы получим:

$$\mathbb{E}_X\,[w(X)] = \int \frac{\sum_i x_i(a\,x_i + \varepsilon_i)}{\sum_i x_i^2} p(\bar{x})p(\bar{\varepsilon})d\bar{x}d\bar{\varepsilon} =$$

$$= a\int \frac{\sum_i x_i^2}{\sum_i x_i^2} p(\bar{x})p(\bar{\varepsilon})d\bar{x}d\bar{\varepsilon} + \int \frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2} p(\bar{x})p(\bar{\varepsilon})d\bar{x}d\bar{\varepsilon}.$$

Мы сократили обозначение для дифференциалов и для плотностей распределений. Первый интеграл равен a (интеграл по всему пространству от плотности распределения). Второй интеграл берется по симметричным относительно нуля интервалам от нечетной по x_i и по ε_i функции, а значит, равен 0. Итак, «средний» коэффициент равен a.

Те же вычисления можно сделать более просто с помощью закона полного математического ожидания $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid Y))$ без интеграла

$$\mathbb{E}_X\left[w(X)\right] = \mathbb{E}_X\left[\frac{\sum_i x_i(a\,x_i + \varepsilon_i)}{\sum_i x_i^2}\right] = a + \mathbb{E}_X\left[\frac{\sum_i x_i\,\varepsilon_i}{\sum_i x_i^2}\right] = a + \mathbb{E}_X\left[\mathbb{E}_\varepsilon\left[\frac{\sum_i x_i\,\varepsilon_i}{\sum_i x_i^2} \mid x\right]\right] =$$
$$= a + \mathbb{E}_X\left[\frac{\sum_i x_i\,\mathbb{E}_\varepsilon\left[\varepsilon_i \mid x\right]}{\sum_i x_i^2}\right] = a + \mathbb{E}_X\left[\frac{\sum_i x_i\,\mathbb{E}_\varepsilon\left[\varepsilon_i\right]}{\sum_i x_i^2}\right] = a.$$

Найдем смещение:

$$\mathbb{E}_x\left[\left(\mathbb{E}_X\left[\mu(X)(x)\right] - \mathbb{E}\left[y|x\right]\right)^2\right] = \mathbb{E}_x\left[\left(a\,x - a\,x\right)^2\right] = 0.$$

Это интуитивно понятный результат: логично, что перебрав все возможные выборки длины ℓ и усреднив по ним значение k, мы обязательно найдем истинную величину коэффициента.

Теперь найдем «среднее» k для произвольной четной $f(x)$. По аналогии с предыдущим случаем:

$$\mathbb{E}_X\left[w(X)\right] = \int \frac{\sum_i (x_i f(x_i))}{\sum_i x_i^2} p(\bar{x})d\bar{x} + \int \frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2} p(\bar{x})p(\bar{\varepsilon})d\bar{x}d\bar{\varepsilon}.$$

Как мы уже выяснили, второй интеграл равен 0. Первый интеграл тоже равен 0, так как подынтегральное выражение — это нечетная по всем x_i функция. Итак, «среднее» значение коэффициента равно нулю. И это тоже понятный результат, потому что четную функцию логично приближать четной, а единственная четная линейная функция, проходящая через 0 — это $y = 0$.

Найдем смещение:

$$\mathbb{E}_x\left[\left(\mathbb{E}_X\left[\mu(X)(x)\right] - \mathbb{E}\left[y|x\right]\right)^2\right] = \mathbb{E}_x\left[\left(0 - f(x)\right)^2\right] = \mathbb{E}_x f^2(x).$$

Чем меньше четная функция $y = f(x)$ похожа на четную линейную $y = 0$, тем больше будет смещение. Таким образом, если мы пытаемся приблизить нелинейную функцию $f(x)$ в классе линейных, мы получаем большое смещение. Обратите внимание, что если бы $f(x)$ не была четной, мы бы не смогли просто аналитически вычислить интегралы.

■

Задача 1.7. *Найдите разброс алгоритма одномерной линейной регрессии для $f(x) = ax$ и для произвольной четной $f(x)$.*

Решение. Для $f(x) = ax$:

$$\mathbb{E}_x\left[\mathbb{E}_X\left[(\mu(X)(x) - \mathbb{E}_X\left[\mu(X)(x)\right])^2\right]\right] = \mathbb{E}_x\left[\mathbb{E}_X\left[\left(ax + \frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}x - ax\right)^2\right]\right] =$$
$$= \left(\mathbb{E}_x x^2\right)\mathbb{E}_X\left[\left(\frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}\right)^2\right] = \sigma_1^2\mathbb{E}_X\left[\left(\frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}\right)^2\right].$$

Математическое ожидание можно немного упростить, раскрыв квадрат суммы в числителе и внеся внутрь суммы математическое ожидание по $\bar{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_\ell)$:

$$\mathbb{E}_X\left[\frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}\right]^2 = \mathbb{E}_{\bar{x}}\left[\frac{\sum_{i\neq j} x_i x_j \mathbb{E}_{\bar{\varepsilon}}\left[\varepsilon_i\varepsilon_j\right] + \sum_i x_i^2\mathbb{E}_{\bar{\varepsilon}}\varepsilon_i^2}{\left(\sum_i x_i^2\right)^2}\right].$$

Так как ε_i и ε_j независимы, $\mathbb{E}\left[\varepsilon_i\varepsilon_j\right] = 0$, а $\mathbb{E}_{\bar{\varepsilon}}\varepsilon_i^2 = \sigma_2^2$. Тогда

$$\mathbb{E}_X\left[\frac{\sum_i x_i\varepsilon_i}{\sum_i x_i^2}\right]^2 = \mathbb{E}_{\bar{x}}\left[\frac{\sum_i x_i^2\sigma_2^2}{\left(\sum_i x_i^2\right)^2}\right] = \sigma_2^2\mathbb{E}_{\bar{x}}\left[\frac{1}{\sum_i x_i^2}\right],$$

а разброс

$$\mathbb{E}_x\left[\mathbb{E}_X\left[(\mu(X)(x) - \mathbb{E}_X\left[\mu(X)(x)\right])^2\right]\right] = \sigma_1^2\sigma_2^2\mathbb{E}_{\bar{x}}\left[\frac{1}{\sum_i x_i^2}\right].$$

Последнее математическое ожидание также можно рассчитать, но мы не будем этого делать. Мы получили, что если шум в ответах небольшой, то и разброс модели будет небольшой.
Для четной $f(x)$:

$$\mathbb{E}_x\left[\mathbb{E}_X\left[(\mu(X)(x) - \mathbb{E}_X\left[\mu(X)(x)\right])^2\right]\right] = \mathbb{E}_x\left[\mathbb{E}_X\left[\left(0 - \frac{\sum_i x_i(f(x_i) + \varepsilon_i)}{\sum_i x_i^2}x\right)^2\right]\right] =$$
$$= \left(\mathbb{E}_x x^2\right)\mathbb{E}_X\left[\left(\frac{\sum_i x_i(f(x_i) + \varepsilon_i)}{\sum_i x_i^2}\right)^2\right] = \sigma_1^2\mathbb{E}_X\left[\left(\frac{\sum_i x_i(f(x_i) + \varepsilon_i)}{\sum_i x_i^2}\right)^2\right].$$

■

Для линейной модели с регуляризатором можно проделать всё то же самое, но интегралы будут более сложными.

§1.3 Разложение для решающих деревьев

Мы выяснили, что линейные модели имеют маленькое смещение, когда истинная зависимость в данных также линейна, и большое смещение, если это не так. С решающими деревьями ситуация противоположна. Мы не будем формально это обосновывать, но интуитивно понятно, что поскольку для любой выборки можно построить дерево, имеющую нулевую ошибку на обучении, то смещение решающего дерева будет небольшое для любой истинной зависимости $f(x)$. Разброс, наоборот, будет большой, потому что при малом изменении в выборке мы можем получить совершенно другое решающее дерево.

С другой стороны, можно ограничивать многообразие деревьев, установив ограничение на глубину или минимальное число объектов в листовых вершинах. Тогда смещение будет увеличиваться, а разброс уменьшаться. В граничном случае, когда $\mu(x) = C = \text{const}$, разброс, очевидно, будет равен 0.

§1.4 Приближенное вычисление интегралов

Разложение на смещение и разброс — это теоретическая конструкция, показывающая, из-за чего происходит переобучние и недообучение алгоритмов. Обычно при анализе сложности алгоритма оперируют терминами «смещение» и «разброс», качественно оценивая их величину (например, мы так сделали с деревьями). Вычислить компоненты аналитически для большинства алгоритмов не представляется возможным. Однако, если есть необходимость количественно оценить их, можно воспользоваться техниками приближенного вычисления интегралов с помощью семплирования.

Если нам нужно оценить математическое ожидание $\mathbb{E}_{x\sim p(x)} f(x) = \int f(x)p(x)dx$, то можно просемплировать выборку $\{x_1, \ldots, x_n\}$ из распределения $p(x)$ и приближенно вычислить интеграл:

$$\mathbb{E}_{x\sim p(x)} f(x) \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

Из областей с большим значением плотности в выборку попадет больше точек, и они внесут больший вклад в значение интеграла. Несложно показать, что данная оценка является несмещенной.

Для вычисления математического ожидания по случайным выборкам нужно сгенерировать несколько выборок:

$$\mathbb{E}_X\left[\mu(X)(x)\right] \approx \frac{1}{n}\sum_{i=1}^{n}\mu(X_i)(x), \quad X_i = \{(x_{i,1}, y_{i,1}), \ldots, (x_{i,\ell}, y_{i,\ell})\},\ x_{i,j}, y_{i,j} \sim p(x, y)$$

2 Композиционные алгоритмы

§2.1 Беггинг

Существуют способы уменьшить разброс алгоритма. Наиболее известный из них — бэггинг. Беггинг заключается в генерации нескольких новых выборок $X_1, \ldots, X_m$ на основе имеющейся, обучении алгоритма на каждой из сгенерированных выборок и усреднении ответов всех алгоритмов на новом объекте.

Задача 2.1. *При бэггинге новую выборку $\widetilde{X}$ составляют, генерируя элементы из X с возвращением. При этом объекты в $\widetilde{X}$ могут повторяться. Будем считать, что число объектов в $\widetilde{X}$ и в X одинаковое и равно ℓ. Найдите вероятность того, что конкретный объект попадет в выборку.*

Решение. Вероятность того, что объект попадет в выборку при одном вытаскивании — $\frac{1}{\ell}$, ℓ — число объектов выборки. Вероятность того, что не попадет — $1 - \frac{1}{\ell}$; не

попадет ни при одном вытаскивании — $(1 - \frac{1}{\ell})^\ell$. Наконец, искомая вероятность

$$\lim_{\ell\to\infty} 1 - \left(1 - \frac{1}{\ell}\right)^\ell = 1 - \frac{1}{e}.$$

■

На лекции было показано, что усреднение с помощью бэггинга уменьшает разброс алгоритма в m раз, если базовые алгоритмы мало коррелированы.

Такой эффект наблюдается при усреднении предсказаний любых алгоритмов регрессии, необязательно полученных бэггингом, если эти алгоритмы выдают слабо коррелированные ответы.

§2.2 Простое голосование

Идею о том, что при выборе итогового предсказания для объекта x с помощью агрегации предсказаний нескольких базовых алгоритмов предсказание будет точнее, можно применить и к классификации. Наиболее простой способ это сделать — построить несколько классификаторов, предсказывающих класс для объекта x, и выбирать класс, который чаще всего предсказывали эти алгоритмы (простое голосование).

Задача 2.2. *Пусть у нас есть три бинарных классификатора, каждый из которых ошибается с вероятностью p. С какой вероятностью будет ошибаться классификатор, построенный с помощью простого голосования? При каких значениях p эта вероятность будет меньше p?*

Решение. Простое голосование выдаст правильный ответ, если не более чем один алгоритм ошибется. Вероятность того, что все три алгоритма ответят правильно, равна $(1-p)^3$, вероятность того, что ровно два из трех ответят правильно: $3p(1-p)^2$. Итоговая вероятность ошибки:

$$1-(1-p)^3-3p(1-p)^2 = 1-1+3p-3p^2+p^3-3p+6p^2-3p^3 = -2p^3+3p^2 = p^2(3-2p).$$

$$-2p^3 + 3p^2 < p;$$
$$p(-2p^2 + 3p - 1) < 0;$$
$$-2p(p-1)(p-\frac{1}{2}) < 0.$$

На участке $p \in [0, 1]$ решением этого неравенства является полуинтервал $p \in (0, \frac{1}{2})$. Таким образом, если каждый из алгоритмов хотя бы чуть лучше, чем случайный, то композиция будет давать меньше ошибок, чем каждый алгоритм по отдельности.

■

Задача 2.3. *Пусть у нас теперь есть n независимых бинарных классификаторов f_i, каждый из которых ошибается с вероятностью $p < \frac{1}{2}$. Для удобства будем рассматривать только нечетные n. Можем ли мы оценить вероятность ошибки классификатора g, построенного с помощью простого голосования?*

Решение. Здесь нам поможет неравенство Чернова. Пусть X_i — индикатор ошибки классификатора f_i. Тогда мы имеем

$$P(\sum_{i=1}^{n} X_i > (1+\delta)np) \leq e^{-\frac{\delta^2}{2+\delta}np}.$$

Так как g построен голосованием, то ошибается он тогда и только тогда, когда $\sum_{i=1}^{n} X_i > \frac{n}{2}$, то есть когда ошиблись более половины классификаторов f_i. Возьмем $\delta_p = \frac{1}{2p} - 1$, для которого верно $(1+\delta_p)np = \frac{n}{2}$. Отсюда

$$P(\sum_{i=1}^{n} X_i > \frac{n}{2}) \leq e^{-\frac{\delta_p^2}{2+\delta_p}np}$$

Тогда вероятность ошибки классификатора g не превосходит e^{-Cn}, где $C = -\frac{\delta_p^2}{2+\delta_p}p$ — некоторая положительная константа. Мы получили, что вероятность ошибки композиции g экспоненциально убывает с ростом n.

■

Задача 2.4. *Рассмотрим задачу классификации с выборкой X и долей объектов с $y = +1$ равной 0.5. Разделим ее на две равные части: X_1 и X_2. Баланс классов в них равен p_1 и p_2 соответственно. Легко показать, что $p_1 = 1 - p_2$.*

Будем говорить, что X_1 и X_2 – это два сегмента нашей выборки. На них используются два разных классификатора $a_1(x)$ и $a_2(x)$. Модель на всей выборке можно записать так:

$$a(x) = a_1(x)\ [x \in X_1] + a_2(x)\ [x \in X_2]$$

Пусть на каждом сегменте работает случайный классификатор:

$$AUC(a_1) = AUC(a_2) = 0.5$$

Может ли $AUC(a)$ быть больше? Каким может быть качество модели на всей выборке?

Решение. Воспользуемся вероятностным определением метрики ROC AUC:

$$\text{AUC}\,(a) = \mathbb{P}\big\{a(x_i) > a(x_j) \mid y_i = +1,\ y_j = -1\big\}$$

Проведем вероятностный эксперимент: выберем случайный объект положительного класса и случайный объект отрицательного класса. Вероятность того, что на первом объекте ответ модели больше, и является ROC AUC модели.

Рассмотрим такой вероятностный эксперимент для классификатора a. Случайный *положительный* объект принадлежит первому сегменту X_1 с вероятностью

$$\frac{p_1}{p_1 + p_2} = \frac{p_1}{p_1 + 1 - p_1} = p_1$$

И с вероятностью $1 - p_1$ – второму сегменту, X_2.

Аналогично случайный *отрицательный* объект принадлежит первому сегменту X_1 с вероятностью

$$\frac{1 - p_1}{1 - p_1 + 1 - p_2} = \frac{1 - p_1}{1 - p_1 + p_1} = 1 - p_1$$

И с вероятностью p_1 – второму сегменту, X_2.

Если и положительный, и отрицательный объект принадлежат X_1, то

$$\mathbb{P}\{a(x_i) > a(x_j) \mid y_i = +1,\ y_j = -1,\ x_i,\ x_j \in X_1\} = \mathrm{AUC}\,(a_1) = 0.5$$

Обозначим эту вероятность auc_{11}. Оба объекта принадлежат первому сегменту с вероятностью $\delta_{11} = p_1(1-p_1)$. Аналогично с вероятностью $\delta_{22} = p_1(1-p_1)$ оба объекта принадлежат X_2, и $auc_{22} = 0.5$.

Но что будет, если положительный объект из X_1, а отрицательный из X_2? Предположим, что предсказания модели на первом сегменте *всегда больше* предсказаний модели на втором.

$$a_1(x_i) > a_2(x_j) \quad \forall x_i \in X_1,\ x_j \in X_2$$

Тогда $\mathrm{auc}_{12} = 1$, и это происходит с вероятностью $\delta_{12} = p_1^2$. А с вероятностью $\delta_{21} = (1-p_1)^2$ получим положительный объект из X_2 и отрицательный – из X_1. В этом случае $\mathrm{auc}_{21} = 0$.

Вычислим $\mathrm{AUC}\,(a)$ по формуле полной вероятности:

$$\mathrm{AUC}\,(a) = \mathrm{auc}_{11}\,\delta_{11} + \mathrm{auc}_{12}\,\delta_{12} + \mathrm{auc}_{21}\,\delta_{21} + \mathrm{auc}_{22}\,\delta_{22} = \\ = 0.5p_1(1-p_1) + p_1^2 + 0.5p_1(1-p_1) = p_1$$

Получаем, что $\mathrm{AUC}\,(a)$ *может приниматъ любые значения от 0 до 1.*

Качество на всей выборке зависит от баланса классов на подвыборках после разбиения на сегменты. Если это разбиение случайно ($p_1 = p_2 = 0.5$), то и качество на всей выборке не будет отличаться от качества на отдельных сегментах. Если же разбиение на сегменты *информативно*, то сам этот факт может усилить модель.

■