

Лекция 7. Ансамбли

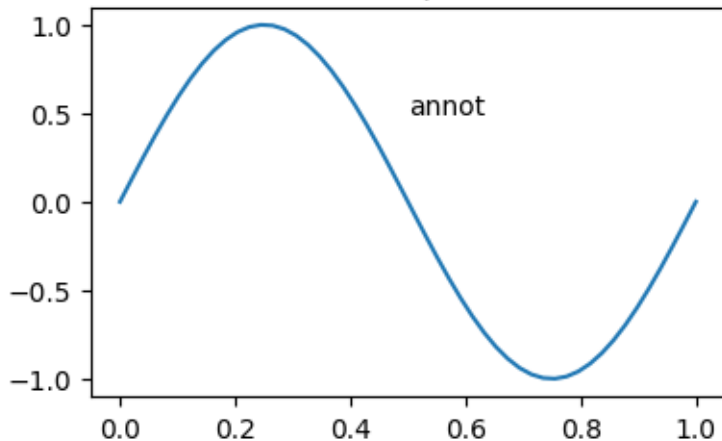
Основы интеллектуального анализа данных

Полузёров Т. Д.

БГУ ФПМИ

TEST

Hello plot



Функционал риска

Рассмотрим задачу регрессии с квадратичной функцией потерь.
Качество алгоритма a :

$$Q(a) = \mathbb{E}_x \mathbb{E}_{X,\varepsilon} [y(x, \varepsilon) - a(x, X)]^2$$

- X — обучающая выборка
- $y = f(x) + \varepsilon$ — целевая зависимость, наблюдаемая с точностью до шума ε
- $a(x, X)$ — ответ алгоритма в точке x , обученного по выборке X
- \mathbb{E}_x — среднее по всем тестовым точкам,
- $\mathbb{E}_{X,\varepsilon}$ — среднее по всем обучающим выборкам и случайному шуму

Смещение и разброс

Ошибку можно представить в виде трех слагаемых:

$$Q(a) = \mathbb{E}_x \text{bias}_X^2(a) + \mathbb{E}_x \text{variance}_X(a) + \sigma^2$$

где

- $\text{bias}_X(a) = f(x) - \mathbb{E}_x[a(x, X)]$ — смещение
- $\text{variance}(a) = \mathbb{E}_X[a(x, X) - \mathbb{E}_x[a(x, X)]]^2$ — разброс
- $\sigma^2 = \mathbb{E}_x \mathbb{E}_\varepsilon[y(x, \varepsilon) - f(x)]^2$ — неустранимый шум

Бутстреп

Пусть имеется выборка X объема ℓ .

С помощью **выбора с возвращением** сформируем новую «выборку» X^1 тоже объема ℓ .

Протредаем эту операцию k раз и получим $\{X^1, \dots, X^k\}$
псевдовыборок.

Такой метод получения псевдовыборок называется **бутстрепои** (bootstrap).

Применяется в статистике для проверки гипотез, построения доверительных интервалов ...

Сгенерируем k псевдовыборок $\{X^1, \dots, X^k\}$ и обучим на каждой из них базовую модель b .

Получим k обученных алгоритмов $b_i(x) = b_i(x, X^i)$, $i = 1, \dots, k$.

Итговый алгоритм есть голосование базовых

$$a(x) = \frac{1}{k} (b_1(x) + \dots + b_k(x))$$

Такой ансамбль алгоритмов — **бэггинг** (bagging, bootstrap aggregation)

Смещение бэггинга

$$\begin{aligned} \text{bias}_X(a) &= f(x) - \mathbb{E}_X[a(x, X)] = \\ &= f(x) - \mathbb{E}_X \left[\sum_{i=1}^k \frac{1}{k} b(x, X^i) \right] = \\ &= f(x) - \mathbb{E}_X b(x, X) = \\ &= \text{bias}_X(b) \end{aligned}$$

Смещение ансамбля определяется смещением базового алгоритма.

Разброс бэггинга

$$\begin{aligned} \text{variance}_X(a) &= \mathbb{E}_X[a(x, X) - \mathbb{E}_x[a(x, X)]]^2 = \\ &= \mathbb{E}_X \left[\frac{1}{k} \sum_{i=1}^k b_i - \mathbb{E}_X \left[\frac{1}{k} \sum_{i=1}^k b_i \right] \right]^2 = \\ &= \frac{1}{k^2} \mathbb{E}_X \left[\sum_{i=1}^k (b_i - \mathbb{E}_X b_i) \right]^2 = \\ &= \frac{1}{k^2} \sum_{i=1}^k \text{variance}_X(b_i) + \frac{1}{k^2} \sum_{i \neq j} \text{cov}(b_i, b_j) \end{aligned}$$

В случае если базовые алгоритмы некоррелированы, то

$$\begin{aligned}\text{variance}_X(a) &= \frac{1}{k^2} \sum_{i=1}^k \text{variance}_X(b_i) = \\ &= \frac{1}{k^2} \sum_{i=1}^k \text{variance}_X(b) = \\ &= \frac{1}{k} \text{variance}_X(b)\end{aligned}$$

Используя некоррелированную композицию алгоритмов, можно добиться уменьшения разброса в k раз!

Случайный лес

На практике строгое выполнение требования некоррелированности — необязательно, достаточно, чтобы базовые алгоритмы были непохожи друг на друга.

Посмотрим бэггинг над решающими деревьями

1 Построение i -го дерева

- ▶ сэмплируется псевдовыборка X^i
- ▶ в процессе построения дерева, в каждой вершине выбирается $n' < n$ признаков и по ним ищется сплит (метод случайных подпространств)

2 итоговый ответ ансамбля: среднее для задач регрессии, наиболее популярный класс для классификации

Случайный лес

Бэггинг над решающими деревьями — случайный лес (random forest). Непохожесть алгоритмов получается за счет обучения на разных псевдовыборках и использовании метода случайных подпространств.

Остаются вопросы

- Какой глубины h строить деревья?
- Сколько признаков n' использовать для обучения?
- Сколько деревьев k использовать в композиции?

Глубина деревьев

Ошибка состоит из смещения и разброса. Разброс снижается за счет композиции, а смещение определяется смещением базового дерева. Поэтому необходимо использовать **деревья с низким смещением**.

- Неглубокие деревья имеют малое число параметров и поэтому строят только верхнеуровневые зависимости. При разных обучающих подвыборках алгоритмы не будут значительно отличаться (низкий разброс, высокое смещение).
- Глубокие деревья наоборот чересчур сильно подстраиваются под данные, поэтому имеют сильный разброс, но низкое смещение.

Используем **глубокие деревья**.

Число признаков

Большое число признаков обеспечивает слабое различие деревьев, поэтому эффект от бэггинга — слабый.

С другой стороны, используя малое число признаков, базовые деревья будут слабыми.

Практическая рекомендация:

- Для задач регрессии — $n' = n/3$
- Для задач классификации — $n' = \sqrt{n}$

Размер ансамбля

С ростом числа деревьев снижается разброс, но число признаков и вариантов подвыборки — ограничены. Поэтому бесконечно уменьшать разброс не получится.

Имеет смысл построить график зависимости ошибки от числа деревьев и остановиться в тот момент, когда ошибка перестанет значительно уменьшаться

Так же ограничением может выступать время работы ансамбля. Большее число деревьев — большее время принятия решения. Однако, алгоритмы стоятся независимо друг от друга и это позволяет распаралеливать построение отдельных деревьев.

Стекинг