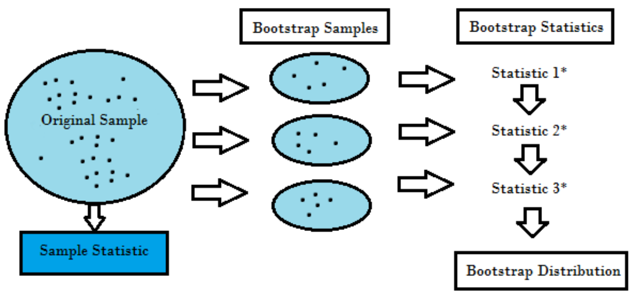
23. Разложение на смещение и разброс

Kazakbaev Rustem

December 2020

1 Идея

Есть обучающая выборка X и есть такая процедура: bootstrap. Из исходной обучающей выборки делаю подвыборку размера l с возвращением.



Делаем n выборок с помощью бутстрэпа: $X_1, X_2, ..., X_n$
Сделаю fit модели и получу базовые алгоритмы: $b_1, b_2, ..., b_n$
Предположу, что знаю распределение на всех возможных объектах: $p(X)$ распределение на X.
$y(X)$ - правильный ответ на объекте X

Посчитаю мат.ожидание ошибки j-ой модели: $E_x(b_j(X) - y(X) = E_x(\epsilon_j^2(X))$. Но это ошибка j-ой модели, теперь попробую найти ошибку на всей выборке: $\frac{1}{N} \cdot E_x(\epsilon_j^2(X))$

1. В среднем у модели ошибка 0: $E_x(\epsilon_j(X)) = 0$

2. Независимость между i и j ошибкой.

Построим теперь новую функцию регрессии, которая будет усреднять ответы построенных нами функций:

$$a(x) = \frac{1}{N} \sum_{j=1}^{N} b_j(x).$$

Найдем ее среднеквадратичную ошибку:
$E_x = (\frac{1}{N} \sum b_j(x) - y(x))^2 = E_x(\frac{1}{N} \epsilon_j)^2 = \frac{1}{N^2} \cdot E_x(\sum \epsilon_j^2 + \underbrace{\sum \epsilon_j \cdot \epsilon_i}_{=0}) = \frac{1}{N^2} \cdot E_x(\sum \epsilon_j^2)$ Будем считать, что все ошибки одинаково распределены, тогда: $\frac{1}{N} \cdot E_X(\epsilon_1^2)$.

Таким образом, усреднение ответов позволило уменьшить средний квадрат ошибки в N раз!

Следует отметить, что рассмотренный нами пример не очень применим на практике, поскольку мы сделали предположение о некоррелированности ошибок, что редко выполняется, хотя это еще может быть возможно.
Предположил, что некореллировано, так как это неправда - делал выборки через бутстрэп.

Если это предположение неверно, то уменьшение ошибки оказывается не таким значительным.

Важный вывод из этой идеи: можно объединять алгоритмы для достижения наилучшего результата!

2 Разложение ошибки на смещение и разброс

Дает понимание, какие модели необходимо брать, чтобы сделать правильные композиции.

Ошибка любой модели складывается из трех факторов:

1. Сложность самой выборки
2. Сходства модели с истинной зависимостью ответов от объектов в выборке
3. Богатство семейства, из которого выбирается конкретная модель

Между этими факторами существует некоторый баланс, и уменьшение одного из них приводит к увеличению другого. Такое разложение ошибки носит название разложения на смещение и разброс.

Есть некоторая обучающая выборка: $X = (x_i, y_i)$. Будем считать, что на пространстве всех объектов и ответов $\times$ существует распределение $p(x, y)$, из которого сгенерирована выборка X и ответы на ней.

Рассмотрим квадратичную функцию потерь

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

Посмотрим на математическое ожидание по x и y квадрата данной ошибки:

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

R(a) называется среднеквадратичным риском. Некоторый способ посчитать ошибку, если знать распределение.

• Разложение ошибки   Bias-Variance Decomposition

→ Модель переобучена?

→ Плохо предсказывает

$$Error = \underbrace{Bias^2(a(x))}_{\text{смещение}} + \underbrace{Var(a(x))}_{\text{разброс}} + \underbrace{\sigma^2}_{\text{шум}}$$

• $Bias(a(x))$ - средняя ошибка по всем возможном наборам данных - смещение

• $Var(a(x))$ - дисперсия ошибки, как сильно различается ошибка при обучении на различных наборах данных - разброс

• σ^2 - неустранимая ошибка (ШУМ)

Доказательство:

• Истинная зависимость $\to y = f(\bar{x}) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

Модель $\to \hat{f}(\bar{x})$

Средняя ошибка на всех об. выборках

$$\bullet Err = E_x\left[(y - \hat{f}(x))^2\right] = E(y^2) + E(\hat{f}(x)^2) - 2\cdot E(y\cdot \hat{f}(x))$$

$$\boxed{E(y^2) = Var(y) + (E(y))^2}$$

$$\Rightarrow Ey^2 = \sigma^2 + f^2$$

$$\Rightarrow E(y\cdot\hat{f}) = E((f+\varepsilon)\cdot\hat{f}) = E(f\hat{f}) + E(\varepsilon\cdot\hat{f}) =$$

$$= f\cdot E(\hat{f}) + \underbrace{E(\overset{0}{\varepsilon})\cdot E(\hat{f})}_{0} = f\cdot E(\hat{f})$$

т.к. Шум в данных не от чего не зависит

$$Var(y) = E(y - E(y))^2 = E(y-f)^2 = \boxed{\sigma^2}$$

$$E(y) = E(f+\varepsilon) = \boxed{E(f)}$$

$$\bullet\ Error = \sigma^2 + f^2 + \underbrace{E(\hat{f}(x))^2}_{2} + 2f\cdot E(\hat{f}) \quad \Big|\quad \underbrace{E(\hat{f}(x))^2}_{2} = Var(\hat{f}) + E\hat{f}^2$$

$$= \sigma^2 + f^2 + Var(\hat{f}) + (E(\hat{f}))^2 - 2\cdot f\cdot E[\hat{f}] =$$

$$= \boxed{(f - E\hat{f})^2 + Var\,\hat{f} + \sigma^2}$$

$$Error = \underbrace{Bias^2(a(x))}_{\text{смещение}} + \underbrace{Var(a(x))}_{\text{разброс}} + \underbrace{\sigma^2}_{\text{шум}}$$

- $Bias(a(x))$ - средняя ошибка по всем возможным наборам данных - смещение
- $Var(a(x))$ - дисперсия ошибки, как сильно различается ошибка при обучении на различных наборах данных - разброс
- σ^2 - неустранимая ошибка (ШУМ)

Совсем плохо

- От чего зависит смещение и разброс?
  - Сложность модели (кол-во параметров)

Сложность↑ => Bias↓ → Var↑

Сложность↓ => Bias↑ → Var↓