Bias – variance разложение

В статистике: $\hat{\theta}$ – оценка θ

$$MSE_{\hat{\theta}}(\theta) := E_\theta(\hat{\theta}-\theta)^2 = \overbrace{(E_\theta\hat{\theta}-\theta)^2}^{\substack{\text{bias}^2 \\ \text{смещение}}} + \overbrace{D_\theta\hat{\theta}}^{\substack{\text{variance} \\ \text{разброс}}}$$

Обозначения: 1) $\hat{y}$ – случайная ф-ция $\mathcal{X} \to \mathcal{Y}$

Случайность может быть обусловлена:

- погрешностью при изм. отклика (случайные y_i)
- случайные признаки
- случайность модели (например, выбор случайного набора пр-ков в вершине)

2) x – неслучайный объект

y – случайный отклик, т.ч. $y = f(x, \varepsilon)$,

где f – детерм ф-ция

ε – случайной шум

$\hat{y}$, ε независимы

$$MSE_{\hat{y}}(x) := E(\hat{y}(x) - y)^2 = E(\hat{y}(x) - f(x,\varepsilon))^2 =$$

$$= E(\hat{y}(x) \pm Ef(x,\varepsilon) - f(x,\varepsilon))^2 =$$

$$= \underbrace{E(\hat{y}(x) - Ef(x,\varepsilon))^2}_{\text{b-v разл. для оценок}} + \underbrace{E(f(x,\varepsilon) - Ef(x,\varepsilon))^2}_{\text{дисперсия } f(x,\varepsilon)} - 2E\overbrace{(\hat{y}(x) - Ef(x,\varepsilon))}^{\text{зависит от } \hat{y}}\overbrace{(f(x,\varepsilon) - Ef(x,\varepsilon))}^{\text{зависит от } \varepsilon} =$$

$$E(f(x,\varepsilon) - Ef(x,\varepsilon)) = 0$$

$$= \underbrace{(E\hat{y}(x) - Ef(x,\varepsilon))^2}_{bias^2} + \underbrace{D\hat{y}(x)}_{variance} + \underbrace{Df(x,\varepsilon)}_{noise}$$

$$E(\hat{y}(x) - y)^2 = (E\hat{y}(x) - Ef(x,\varepsilon))^2 + D\hat{y}(x) + Df(x,\varepsilon)$$

Следствие 1

Пусть шум аддитивен и несмещ. $f(x, \varepsilon) = h(x) + \varepsilon$

$E\varepsilon = 0 \qquad D\varepsilon = \sigma^2$

Тогда $E(\hat{y}(x) - y)^2 = (E\hat{y}(x) - h(x))^2 + D\hat{y}(x) + \sigma^2$

Следствие 2

Пусть X — новый объект, но теперь он случаен

Тогда $E(\hat{y}(X) - y)^2 = E\Big[\underbrace{E\big((\hat{y}(X) - y)^2 \mid X\big)}_{\substack{X - \text{фикс.} \Rightarrow \\ \text{уже посчитано}}}\Big] =$

УD

$$D(X|Y) = E(X^2|Y) - E(X|Y)^2$$

УМО

$$E(X|Y) = f(Y)$$

$$E(X|Y=y) = f(y)$$

$$EE(X|Y) = EX$$

$$E(\hat{y}(X) - y)^2 = \Big(E(\hat{y}(X)|X) - E(f(X,\varepsilon)|X)\Big)^2 + ED(\hat{y}(X)|X) + ED(f(X,\varepsilon)|X)$$

Ridge – регрессия

Применим разложение для аддитивного несмещ. шума.

$$\hat{y}(x) = x^T \hat{\Theta} \qquad h(x) = x^T \Theta \qquad y = X\Theta + \varepsilon$$

$$E\varepsilon = 0$$

$$\hat{\Theta} = (X^T X + \lambda I_d)^{-1} X^T y \qquad D\varepsilon = \sigma^2 I_n$$

$$E\,\hat{y}(x) = x^T E\hat{\Theta} = x^T (X^T X + \lambda I_d)^{-1} X^T X \Theta$$

$$D\,\hat{y}(x) = x^T D\hat{\Theta}\, x = x^T (X^T X + \lambda I_d)^{-1} X^T X (X^T X + \lambda I_d)^{-1} x\, \sigma^2$$

$$bias^2 = (E\hat{y}(x) - h(x))^2 = \ldots$$

$$variance = D\hat{y}(x) = \ldots$$

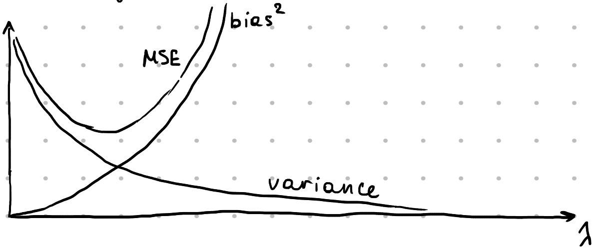
МНК ($\lambda = 0$)

$bias^2 = 0$

$variance = x^T (X^T X)^{-1} x \, \sigma^2$

Вывод: $X^T X$ близка к вырожденной $\Rightarrow$ в МНК разброс большой

Ridge - var < МНК - var



Смесь моделей

$$\boxed{\operatorname{corr}(x,y) = \frac{\operatorname{cov}(x,y)}{\sqrt{Dx\, Dy}}}$$

$$\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t(x)$$

~~$\hat{y}_t$ одинаково распределены~~ $\operatorname{cov}(\hat{y}_a, \hat{y}_b) = \operatorname{cov}(\hat{y}_c, y_d)$

$$bias^2 = \left(E\,\hat{y}(x) - h(x)\right)^2 = \left(E\,\hat{y}_1(x) - h(x)\right)^2$$

т.е. смещение композиции = смещ. одной модели

variance $\quad D\hat{y}(x) = \frac{1}{T^2}\sum_{t_1=1}^{T}\sum_{t_2=1}^{T} \operatorname{cov}\left(\hat{y}_{t_1}(x), \hat{y}_{t_2}(x)\right) =$

$$= \frac{1}{T} D\,\hat{y}_1(x) + \frac{T-1}{T} \operatorname{cov}\left(\hat{y}_1(x), \hat{y}_2(x)\right) =$$

$$\underset{\parallel}{} \quad D\hat{y}_1(x) \; \operatorname{corr}\left(\hat{y}_1(x), \hat{y}_2(x)\right)$$

$$= D\hat{y}_1(x)\left[\frac{1}{T} + \frac{T-1}{T}\operatorname{corr}\left(\hat{y}_1(x), \hat{y}_2(x)\right)\right]$$

Вывод: Разброс тем меньше, чем меньше корреляция моделей.

Bagging

Композиция вида $\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} \hat{y}_t(x)$

где $\hat{y}_t$ построена по случ. подвыборке обучающей выборки с возвращениями.

Замечание: $\hat{y}_t$ могут быть из разных семейств (лин / дерево / ...)

Частный случай — Random Forest

Bagging:

$$\left.\begin{array}{ll} \text{bias} & \text{малое} \\ \text{varianse} & \text{одной модели большой} \\ \text{corr} & \text{моделей низкая} \end{array}\right\} \Rightarrow \begin{array}{ll} \text{bias} & \text{маленький} \\ \text{variance} & \text{маленький} \end{array}$$

Как получить модели с малой корреляцией?

- разные типы моделей, разные гиперпараметры
- обучать на разных признаках
- разная предобработка данных