Johns Hopkins University
EN. 601.475 Machine Learning

# Bias-Variance & Linear Regression

Xuan Zhang
Sep. 9. 2022

I. Empirical Risk Minimization (ERM)

II. Bias-Variance decomposition

III. Linear Regression

VI. MLE & MAP

I.ERM

# Empirical Risk Minimization

## Population Risk / Expected Risk

$$R(f) = \mathbb{E}_{XY}[\text{loss}(Y, f(X))]$$

## Bayes Optimal Rule

$$f^* = \underset{f}{\text{argmin}}\, R(f)$$

## Empirical Risk

$$R(f, D) = \frac{1}{N}\sum_{n=1}^{N} \text{loss}(y_n, f(x_n))$$

## Empirical Risk Minimization

$$\hat{f}_{ERM} = \underset{f}{\text{argmin}}\, R(f, D)$$

# Why does this approximation work?

Law of large numbers: sample average converges to the expected value as sample size approaches $\infty$.

$\bar{X}_N \to \mu$ as $N \to \infty$.

$$\frac{1}{n}\sum_{n=1}^{N} loss(y_n, f(x_n)) \xrightarrow[\text{numbers}]{\text{Law of Large}} \mathbb{E}_{XY}[loss(Y, f(X))]$$

$$= \int loss(Y, f(x))\, p(x,y)\, dx, y$$

# Excess Error Decomposition

$\mathcal{H}$: our hypothesis space

$R(f)$: expected risk / population risk

$R(f, D)$: empirical risk

$f^{**} = \arg\min_{f} R(f)$ : the function that achieves the minimal possible population risk.

$f^{*} = \arg\min_{f \in \mathcal{H}} R(f)$ : the function that achieves the minimal possible population risk in our hypothesis space.

$f^{*}_{N} = \arg\min_{f \in \mathcal{H}} R(f, D)$ : the function that achieves the minimal empirical risk in our hypothesis space.

excess error:

$$\mathbb{E}[R(f^{*}_{N}) - R(f^{**})] = \underbrace{\boxed{\mathbb{E}[R(f^{*}) - R(f^{**})]}}_{\text{approximation error}} + \underbrace{\boxed{\mathbb{E}[R(f^{*}_{N}) - R(f^{*})]}}_{\text{estimation error / generation error}}$$

# Excess Error Decomposition

$$\mathbb{E}[R(f_N^*) - R(f^{**})] = \underbrace{\boxed{\mathbb{E}[R(f^*) - R(f^{**})]}}_{\text{approximation error}} + \underbrace{\boxed{\mathbb{E}[R(f_N^*) - R(f^*)]}}_{\text{estimation error / generation error}}$$

approximation error: ( determined by the capacity of $\mathcal{H}$ )
measures how closely our hypothesis space $\mathcal{H}$ can model the true optimal function $f^{**}$.

generation error: ⟨determined by $N$ & the capacity of $\mathcal{H}$⟩
measures the difference in estimated risk due to having a finite training set.

# II. Bias-Variance Decomposition

# Expected Squared Loss Decomposition

Within a hypothesis space $\mathcal{H}$, e.g. the space of regression functions,

$f(x)$ : prediction function

$y$ : provided label (noisy)

$h(x)$ : true prediction function, true label $\quad h(x) = \mathbb{E}[y|x] = \int y\,p(y|x)\,dy$

expected squared loss:

$$\mathbb{E}[L] = \mathbb{E}[(f(x) - y)^2]$$

error from $\quad$ model $\quad$ data

$$= \boxed{\mathbb{E}[(f(x) - h(x))^2]} + \boxed{\mathbb{E}[(h(x) - y)^2]}$$

depends on the choice for $f(x)$. we want to find a $f(x)$ to minimize this term.

arises from the intrinsic noise on the data

# Expected Squared Loss Decomposition

## (Derivation)

$$\mathbb{E}[L] = \mathbb{E}[(f(x) - y)^2]$$

$$= \mathbb{E}[(f(x) - h(x) + h(x) - y)^2]$$

$$= \mathbb{E}[(f(x) - h(x))^2] + \mathbb{E}[(h(x) - y)^2]$$

$$+ 2\underbrace{\mathbb{E}[(f(x) - h(x))(h(x) - y)]}_{①}$$

$$① = \mathbb{E}_{XY}[(f(x) - \mathbb{E}[y|x])(\mathbb{E}[y|x] - y)]$$

$$= \mathbb{E}_{X}[(f(x) - \mathbb{E}[y|x])\underbrace{\mathbb{E}_{Y|X}[\mathbb{E}[y|x] - y]}_{②}]$$

$$② = \mathbb{E}_{Y|X}[y|x] - \mathbb{E}_{Y|X}[y] = 0$$

$$\Downarrow \qquad \qquad \Downarrow$$

$$\int y\, p(y|x)\, dy \overset{\text{same}}{=\!=} \int y\, p(y|x)\, dy$$

# Bias - Variance Decomposition

expected squared loss:

$$\mathbb{E}[L] = \mathbb{E}[(f(x) - y)^2]$$

error from model data

$$= \boxed{\mathbb{E}[(f(x) - h(x))^2]} + \boxed{\mathbb{E}[(h(x) - y)^2]}$$

depends on the choice for $f(x)$. we want to find a $f(x)$ to minimize this term.

arises from the intrinsic noise on the data

---

first item depends on dataset $D$:

$$\underbrace{\boxed{\mathbb{E}_D[(f(x;D) - h(x))^2]}}_{\text{model risk}} = \underbrace{\boxed{(\mathbb{E}_D[f(x;D)] - h(x))^2}}_{(\text{bias})^2} + \underbrace{\boxed{\mathbb{E}_D[(f(x;D) - \mathbb{E}_D[f(x;D)])^2]}}_{\text{variance}}$$

expected squared loss $= \underbrace{\boxed{(\text{bias})^2 + \text{variance}}}_{\text{model risk}} + \text{noise}$

# Bias - Variance Decomposition (Derivation)

model risk

$$\boxed{\mathbb{E}_D[(f(x;D)-h(x))^2]}$$

substract and add $\mathbb{E}_D[f(x;D)]$

$$= \mathbb{E}_D[(f(x;D)-\mathbb{E}_D[f(x;D)]+\mathbb{E}_D[f(x;D)]-h(x))^2]$$

variance

$$= \boxed{\mathbb{E}_D[(f(x;D)-\mathbb{E}_D[f(x;D)])^2]} + \boxed{\mathbb{E}_D[(\mathbb{E}_D[f(x;D)]-h(x))^2]} \; ①$$

$$+ \underbrace{2\mathbb{E}_D[\{f(x;D)-\mathbb{E}_D[f(x;D)]\}\{\mathbb{E}_D[f(x;D)]-h(x)\}]}_{=0}$$

$$① \quad = \boxed{(\mathbb{E}_D[f(x;D)]-h(x))^2}$$

$(\text{bias})^2$

# Bias - Variance Tradeoff

$$\underbrace{\boxed{\mathbb{E}_D[(f(x;D)-h(x))^2]}}_{\text{model risk}} = \underbrace{\boxed{(\mathbb{E}_D[f(x;D)]-h(x))^2}}_{(\text{bias})^2} + \underbrace{\boxed{\mathbb{E}_D[(f(x;D)-\mathbb{E}_D[f(x;D)])^2]}}_{\text{variance}}$$

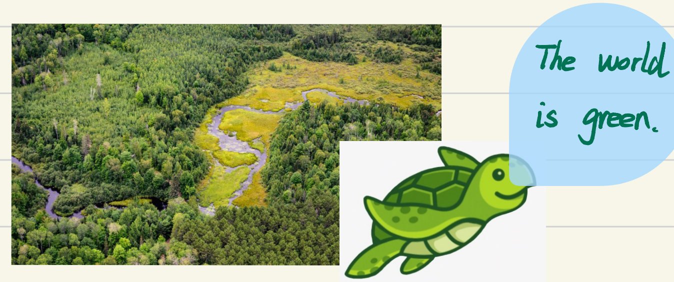
bias : represents the extent to which the average prediction over all data sets differs from the best prediction function.

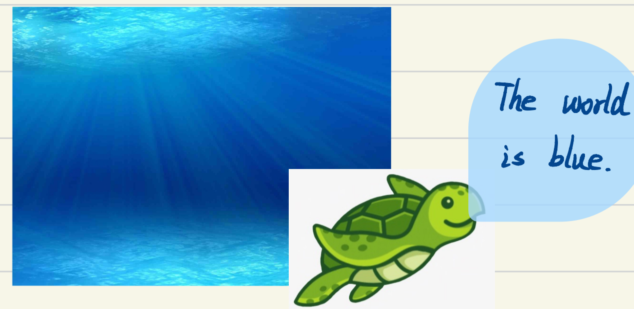
variance: measures the extent to which the solutions for individual data sets vary around the average.

i.e. measures the extent to which the function $f(x;D)$ is sensitive to the particular choice of data set.

# Bias - Variance Tradeoff


high bias (underfitting)

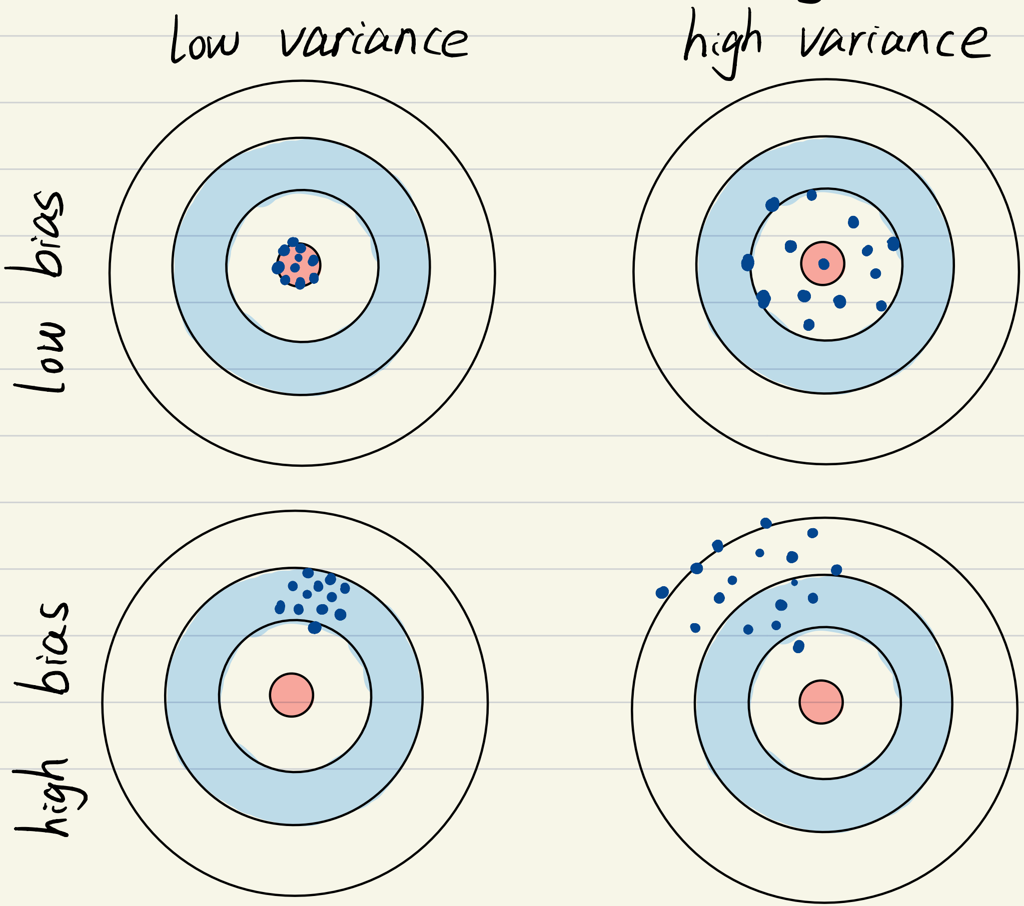



high variance (overfitting)

* Skate is the only animal that has been confirmed to see only in black and white.
* That turtle is purely made up.

# Bias - Variance Tradeoff



# Ⅲ. Linear Regression

# Linear Regression

$x_i \in \mathbb{R}^D$

$y \in \mathbb{R}^N$

$w \in \mathbb{R}^D$

$$p(y|x,\theta) = \mathcal{N}(y \mid w^T x, \sigma^2)$$

Likelihood $\prod_{i=1}^{N} \mathcal{N}(y_i \mid w^T x_i, \sigma^2)$

log likelihood $\sum_{i=1}^{N} \log \left[ \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right) \right]$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - w^T x_i)^2 - \frac{N}{2} \log(2\pi\sigma^2)$$

maximize log likelihood $\iff$ minimize $\frac{1}{2} \sum_{i=1}^{N} (y_i - w^T x_i)^2$

# Least Squares Estimation

residual sum of squares (RSS):

$X \in \mathbb{R}^{N \times D}$

$$RSS(w) = \frac{1}{2}\sum_{i=1}^{N}(y_i - w^T x_i)^2$$

$$= \frac{1}{2}\|Xw - y\|_2^2$$

$$= \frac{1}{2}(Xw-y)^T(Xw-y)$$

Recall 2-norm (Euclidean norm)

$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$ note that $\|x\|_2^2 = x^T x$

optimize:

$$\nabla_w RSS(w) = X^T X w - X^T y = 0$$

$$X^T X w = X^T y$$

least squares solution: $\boxed{w = (X^T X)^{-1} X^T y}$

# Least Squares Estimation (Derivation)

$$RSS(w) = \frac{1}{2}(Xw - y)^T(Xw - y)$$

$$= \frac{1}{2} w^T X^T X w - \frac{1}{2} y^T X w - \frac{1}{2} w^T X^T y + \frac{1}{2} y^T y$$

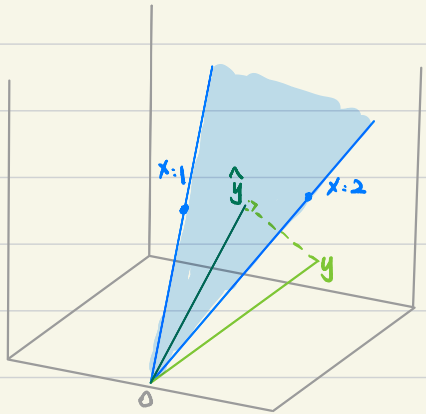
$$\nabla_w RSS(w) = \frac{1}{2} \frac{\partial w^T X^T X w}{\partial w} ① - \frac{1}{2} \frac{\partial y^T X w}{\partial w} ② - \frac{1}{2} \frac{\partial w^T X^T y}{\partial w} ③$$

$$= \frac{1}{2}(X^T X + X^T X) w - \frac{1}{2} X^T y - \frac{1}{2} X^T y$$

$$= X^T X w - X^T y$$

Vector derivatives: ① $\frac{\partial x^T A x}{\partial x} = (A + A^T) x$ ② $\frac{\partial a^T x}{\partial x} = a$ ③ $\frac{\partial x^T a}{\partial x} = a$

# Geometric interpretation of least squares




$\hat{y}$ lies in the linear subspace spanned by $X$: $\hat{y} = Xw$

we want to find

$$\hat{y}^* = \underset{\hat{y} \in \mathrm{span}(\{x_{:1}, \cdots x_{:d}\})}{\arg\min} \|y - \hat{y}\|_2$$

$x_{:,d}$ : $d$th column of $X$

To minimize $\|y - \hat{y}\|_2$, we want $(y - \hat{y})$ to be orthogonal to every column of $X$:

$$X^T(y - Xw) = 0 \Rightarrow w = (X^T X)^{-1} X^T y$$

# VI. MLE & MAP

# MLE

maximum likelihood estimation (MLE):

Pick the parameters that assign the highest probability to data.

$$\boxed{\hat{\theta}_{mle} \triangleq \underset{\theta}{\operatorname{argmax}}\, p(D \mid \theta)}$$

likelihood $p(D \mid \theta) = \prod_{i=1}^{N} p(y_i \mid x_i, \theta)$ iid assumption

log likelihood $LL(\theta) \triangleq \log p(D \mid \theta) = \sum_{i=1}^{N} \log p(y_i \mid x_i, \theta)$

negative log likelihood $NLL(\theta) \triangleq -\log p(D \mid \theta) = -\sum_{i=1}^{N} \log p(y_i \mid x_i, \theta)$

# MAP

maximum a posterior estimation (MAP):

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \log p(\theta \mid D)$$

$$= \underset{\theta}{\operatorname{argmax}} [\log p(D \mid \theta) + \underbrace{\log p(\theta)}_{\text{prior}} - \text{const}]$$

* Will learn more about MAP when we learn Regularization.

Questions?