UNITS 1-2

$$y \approx \beta_0 + \beta_1 x_1 + \dots + \beta_d x_d \triangleq \hat{y}$$

training :
$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\underline{y}} \approx \underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}}_{\underline{A}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}}_{\underline{\beta}} \triangleq \underbrace{\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}}_{\underline{\hat{y}}}$$

test : $\hat{y}(\underline{x}) = [1 \quad \overset{\text{test}}{\underline{x}^T}]\, \underline{\hat{\beta}} \Leftarrow$ learned

$$RSS(\underline{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \| \underline{y} - \underline{A}\underline{\beta} \|^2$$ "least squares"

to find $\underline{\hat{\beta}}_{LS}$: $\frac{\partial}{\partial \beta_j} RSS(\underline{\beta}) \Big|_{\hat{\beta}_j} = 0 \ \forall j \Rightarrow \underline{\hat{\beta}}_{LS} = (\underline{A}^T \underline{A})^{-1} \underline{A}^T \underline{y}$

R^2 coefficient of determination : $R^2 \triangleq 1 - \frac{RSS}{n s_y^2}$

- $R^2 = 1$: perfect
- $R^2 = 0$: trivial
- $R^2 < 0$: impossible on training data for $\underline{\hat{\beta}}_{LS}$

categorical features $x_j \in \{A, B, C\}$

- use one-hot coding : turn x_j into a one-hot binary vector
- coding the intercept : each category gets a unique intercept β_0
- code the slope : " " " " " slope

nonlinear transformations $\longrightarrow$ new features

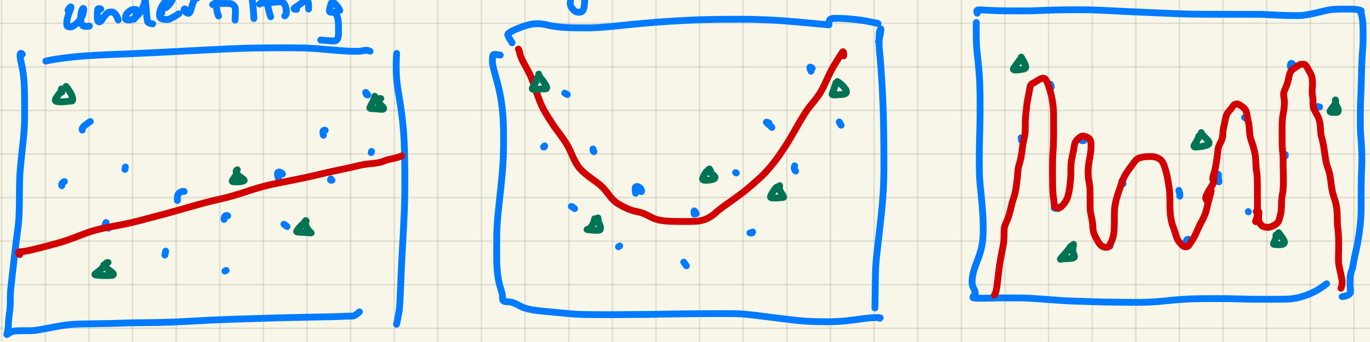
e.g. polynomial regression

$$y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_d x^d$$ $\leftarrow$ use $x_j = x^j$

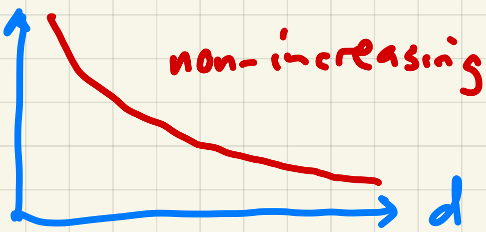
but how to choose d ?

UNIT 3

underfitting | good fit | overfitting



training RSS (or MSE)



: suggests to use most complicated model $\Rightarrow$ overfitting

solution: cross-validation = choose model order using samples different from ~~training samples~~ ones used to get $\hat{\beta}$

options: 1) test / train split

2) K-fold: train on K-1 folds, test on remaining fold } repeat for all K combinations

$$\overline{RSS}_d = \frac{1}{K}\sum_{k=1}^{K} RSS_{k,d} \quad \leftarrow \text{estimate of } E\{RSS_d\}$$

$\uparrow$ random

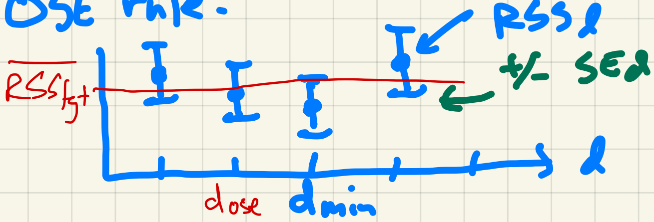
simple approach: choose $d_{min} = \arg\min_d \overline{RSS}_d$

"standard error" on $\overline{RSS}_d$

$$SE_d \triangleq \frac{\hat{\sigma}}{\sqrt{K}} \quad \text{where} \quad \hat{\sigma} = \sqrt{\frac{1}{K-1}\sum_{i=1}^{K}\left(RSS_{i,d} - \overline{RSS}_d\right)^2}$$

$\hookrightarrow$ estimate of $\sqrt{\text{var}\{RSS_d\}}$

OSE rule:



$$\overline{RSS}_{tgt} = \overline{RSS}_{d_{min}} + SE_{d_{min}}$$

$$d_{ose} \triangleq \min_d \{ d : \overline{RSS}_d \leq \overline{RSS}_{tgt} \}$$

$d \leftarrow$ simplest

Bias - Variance tradeoff

- true model : training: $y_i = f(\underline{x}_i) + \varepsilon_i$, testing: $y = f(\underline{x}) + \varepsilon$ } ε_i and ε are iid, zero-mean, variance σ^2

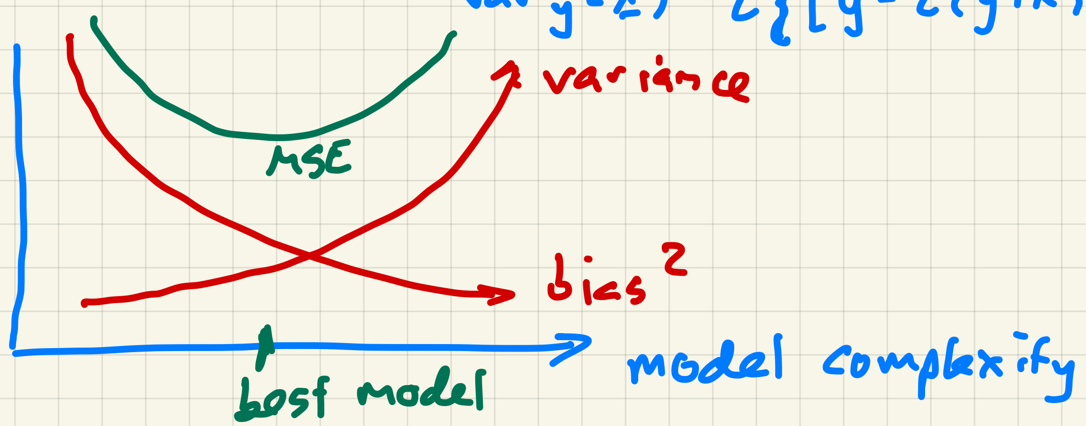
- prediction model : $\hat{y} = \hat{f}(\underline{x}, \hat{\underline{\beta}})$ ← random trained parameters depend on random $\underline{x}_i$ & ε_i

- metric : $MSE_{\hat{y}}(\underline{x}) \triangleq E\{(y-\hat{y})^2 \mid \underline{x}\}$

we derived $MSE_{\hat{y}}(\underline{x}) = \sigma^2 + [bias_{\hat{y}}(\underline{x})]^2 + var_{\hat{y}}(\underline{x})$

where $bias_{\hat{y}}(\underline{x}) = E\{\hat{y} - y \mid \underline{x}\}$

$var_{\hat{y}}(\underline{x}) = E\{[\hat{y} - E\{\hat{y} \mid x\}]^2 \mid \underline{x}\}$



- special case : LS linear regression

$d < d_{true}$: bias $\neq 0$ ⇒ underfitting

$d \geq d_{true}$: bias $= 0$

$$E\{var_{\hat{y}}(\underline{x})\} = \frac{d+1}{n}\sigma^2$$

UNIT 4

Feature Selection: choose best subset of d features

- exhaustive search: optimal, but complexity grows as 2^d
- stepwise regression: greedy, but useful ... complexity d^2
- ranking based on univariate statistics (correlation)
- regularization-based methods

① LASSO: $\arg\min_{\underline{\beta}} \{ \|\underline{y} - \underline{X}\underline{\beta}\|_2^2 + \overbrace{\alpha \|\underline{\beta}\|_1}^{\text{L1 regularization}} \}$

- sets a subset of $\{\beta_j\}$ to zero, shrinks remaining β_j
- we use the indices of nonzero $\{\beta_j\}$ but not their values

fit a LS model!

② Ridge: $\arg\min_{\underline{\beta}} \{ \|\underline{y} - \underline{X}\underline{\beta}\|_2^2 + \overbrace{\alpha \|\underline{\beta}\|_2^2}^{\text{L2 regularization}} \}$

- not useful for feature selection (no zero-valued β_j)
- useful with correlated features

Probabilistic Interpretations

① Maximum Likelihood (ML): $\hat{\underline{\beta}}_{ML} \triangleq \arg\max_{\underline{\beta}} \overbrace{p(\underline{y} \mid \underline{X}, \underline{\beta})}^{\text{likelihood fn}}$ ($\underline{y}$, $\underline{X}$ fixed)

$= \arg\min_{\underline{\beta}} [-\ln p(\underline{y} \mid \underline{X}, \underline{\beta})]$

- ML estimation of $\underline{\beta}$ under $\underline{y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$, $\underline{\varepsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I)$ gives LS estimation: $\hat{\underline{\beta}}_{ML} = \arg\min_{\underline{\beta}} \|\underline{y} - \underline{X}\underline{\beta}\|_2^2$

② MAP: $\hat{\underline{\beta}}_{MAP} = \arg\max_{\underline{\beta}} p(\underline{\beta} \mid \underline{X}, \underline{y})$

$= \arg\min_{\underline{\beta}} [-\ln p(\underline{\beta} \mid \underline{X}, \underline{y})]$

$p(\underline{\beta} \mid \underline{X}, \underline{y}) = \frac{p(\underline{y} \mid \underline{X}, \underline{\beta})\, p(\underline{\beta})}{\cancel{p(\underline{y} \mid \underline{X})}}$

- Ridge is MAP under $\beta_j \sim \mathcal{N}(0, v)$
- LASSO is MAP under $\beta_j \sim \text{Laplacian}(v)$

} and linear/Gaussian likelihood