are in majority.

Bias–Variance Decomposition of Error

Let the test example $z_1\, z_2\, z_3 \cdots z_t$ be fixed.

Variance in the error reflects the instability in classifying a given test example z_i

Assume that the true label y_i for an example x_i be given as

Ground truth $y_i = \underbrace{f(x_i)}\, + \underbrace{\epsilon_i}$ noise — measurement zero mean

The true function f is unknown.

Assume that a model learned using the training set S predicts the label of a test example z_i as

prediction $\hat{y}_i = g(z_i, S)$ Model depends on S.

The mean square error (MSE) of the prediction over a test dataset is given as

t: test example $\frac{1}{t} \sum_{i=1}^{t} \left(\hat{y}_i - y_i\right)^2$ Pred GT

$$= \frac{1}{t} \sum_{i=1}^{t} \Big(\underbrace{g(z_i, S)}_{\hat{y}_i} - \underbrace{f(z_i) - \epsilon_i}_{y_i}\Big)^2$$

Expected MSE over different choices of the training set on the same test examples.

$$\underset{S}{\mathbb{E}}\left[MSE\right] = \frac{1}{t} \sum_{i=1}^{t} \underset{S}{\mathbb{E}}\left[\Big(\underbrace{g(z_i, S) - f(z_i)}_{a} - \underbrace{\epsilon_i}_{b}\Big)^2\right]$$

$$\mathbb{E}[a^2+b^2+2ab] = \mathbb{E}[a^2]+\mathbb{E}[b^2]+\mathbb{E}(2ab)$$

$$\mathbb{E}[MSE] = \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\Big[\underbrace{(g(\underline{z}_i,S)-f(\underline{z}_i))^2}_{a^2}\Big] + \underbrace{\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}[\underbrace{\epsilon_i^2}_{b^2}]}_{B.}$$

2nd term variance of the noise

$$+ \frac{2}{t}\sum_{i=1}^{t}\underbrace{\mathbb{E}\big[g(\underline{z}_i,S)-f(\underline{z}_i)\big]\,\mathbb{E}[\epsilon_i]}_{a\cdot b}$$

mean of the noise is zero

First term: $\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\Big[\big(\underbrace{g(\underline{z}_i,S)}_{\text{prediction}}-\underbrace{f(\underline{z}_i)}_{\text{true function}}\big)^2\Big]$

$$= \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\Big[\Big\{\underbrace{\big(f(\underline{z}_i)-\mathbb{E}_S[g(\underline{z}_i,S)]\big)}_{a} + \underbrace{\big(\mathbb{E}_S[g(\underline{z}_i,S)] - g(\underline{z}_i,S)\big)}_{b}\Big\}^2\Big]$$

$$= \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_S\Big[\underbrace{\big(f(\underline{z}_i)-\mathbb{E}[g(\underline{z}_i,S)]\big)^2}_{a^2}\Big]$$

A(I)

$$+ \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}_S\Big[\underbrace{\big(\mathbb{E}[g(\underline{z}_i,S)] - g(\underline{z}_i,S)\big)^2}_{b^2}\Big]$$

A(II)

$$+ \frac{2}{t}\sum_{i=1}^{t}\mathbb{E}_S\Big[\underbrace{\big(f(\underline{z}_i)-\mathbb{E}[g(\underline{z}_i,S)]\big)}_{a}\Big] \times \underbrace{\mathbb{E}_S\Big[\big(\mathbb{E}_S[g(\underline{z}_i,S)] - g(\underline{z}_i,S)\big)\Big]}_{b.\ =0}$$

$$\mathbb{E}_S\mathbb{E}_S[g(z_i,S)] = \mathbb{E}_S[g(z_i,S)] \qquad -\mathbb{E}_S[g(\underline{z}_i,S)]$$

$$\mathbb{E}_{S}\left[MSE\right] = \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\left(g(z_i,S)-f(z_i)\right)^2\right] + \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\epsilon_i^2\right]$$

$$= \underline{\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\left(f(z_i)-\mathbb{E}\left[g(z_i,S)\right]\right)^2\right]}$$

$$+ \underline{\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\left(\mathbb{E}_{S}\left[g(z_i,S)\right]-g(z_i,S)\right)^2\right]} + \underline{\frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\epsilon_i^2\right]}$$

bias^2

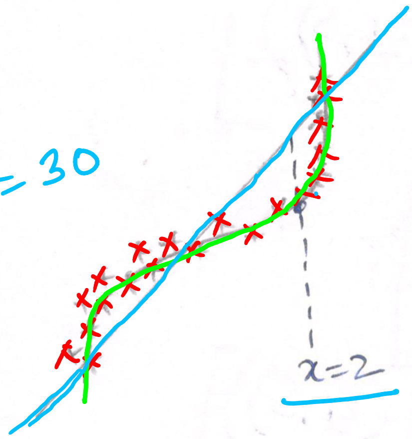
Variance of noise

$$\mathbb{E}\left[MSE\right] = \frac{1}{t}\sum_{i=1}^{t}\left[\left(\underbrace{f(z_i)}_{\text{true function target}} - \underbrace{\overbrace{\mathbb{E}_{S}\left[g(z_i,S)\right]}^{\text{average prediction}}}\right)^2\right]$$

$$+ \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\left(\underbrace{g(z_i,S) - \mathbb{E}_{S}\left[g(z_i,S)\right]}_{\text{Variance of the model prediction}}\right)^2\right] + \frac{1}{t}\sum_{i=1}^{t}\mathbb{E}\left[\epsilon_i^2\right]$$

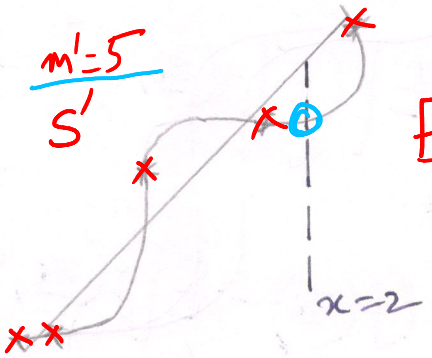
$m = 30$

S

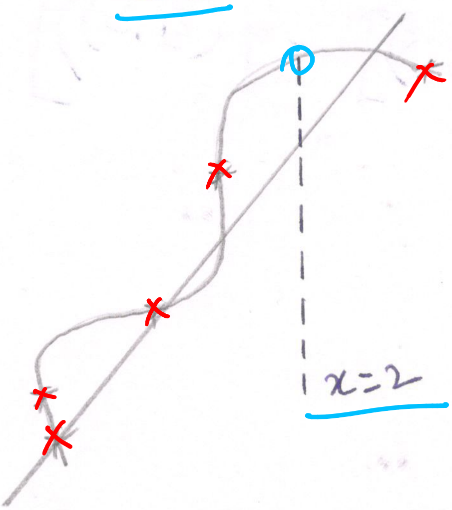


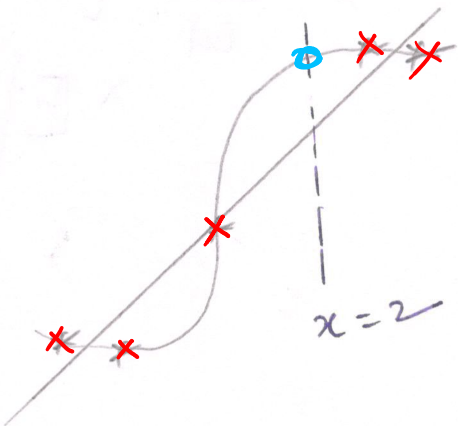
$m' = 5$

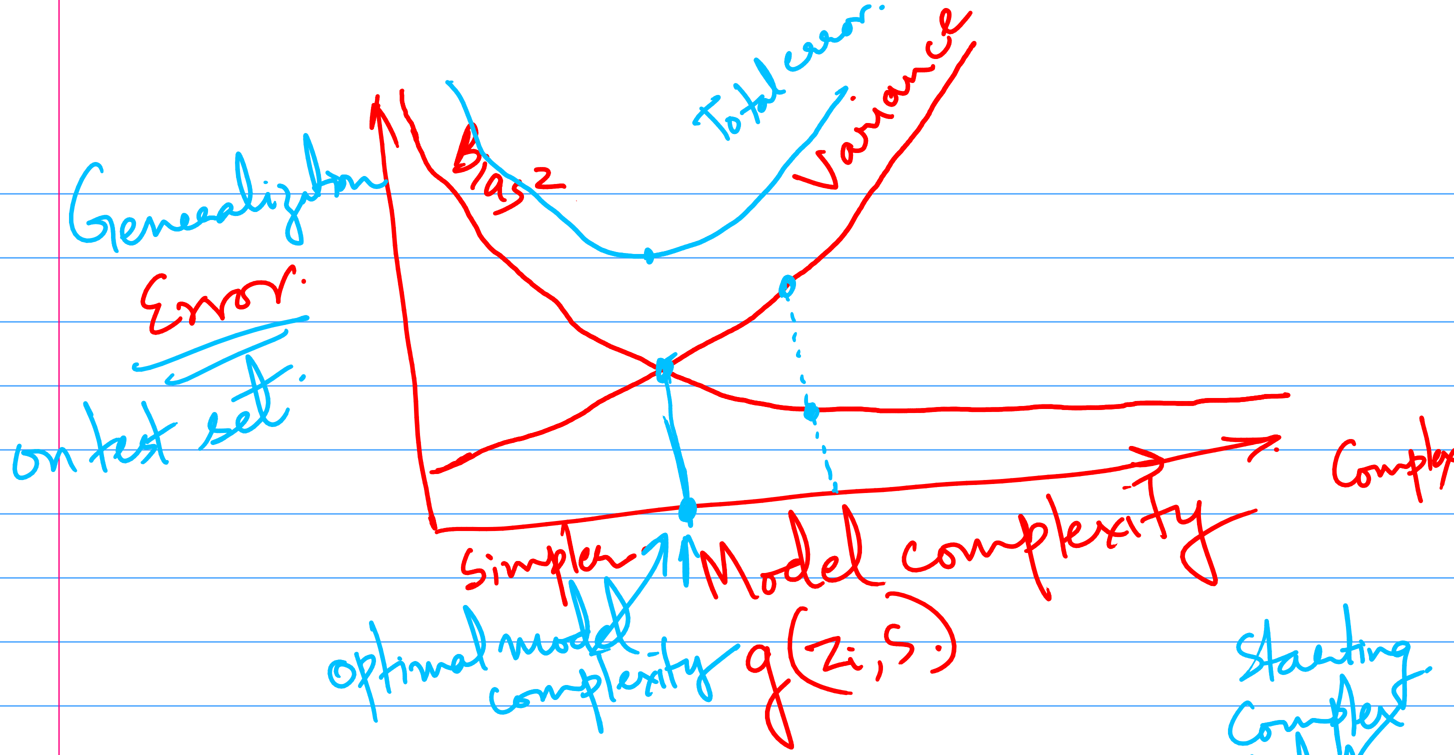
S'



Error = Bias^2 + Variance + Noise







Decision trees] Pruning
Early stopping

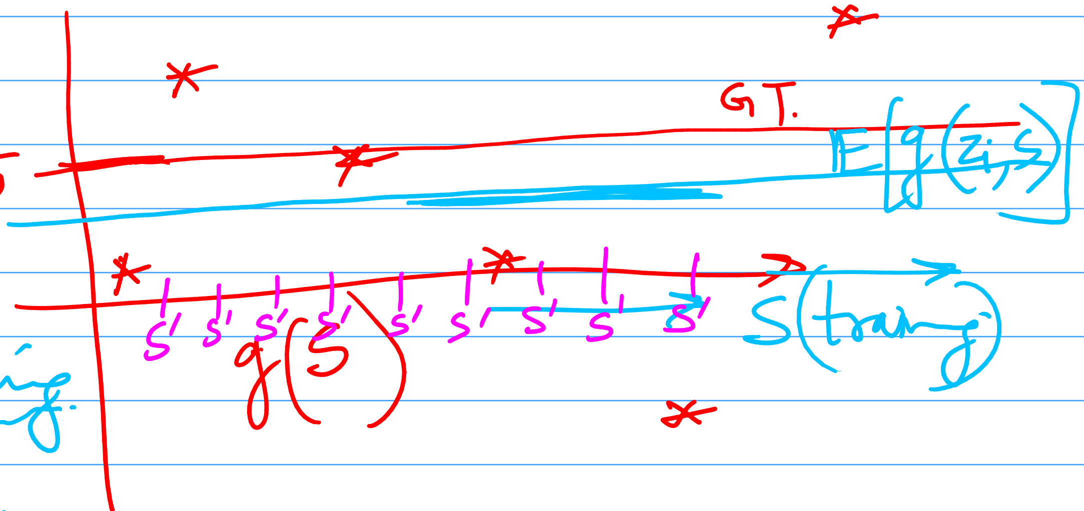
SVMs
λ | C.

Starting Complex Model

Final model of the right complexity.

Regularization

High Variance
the classifier is sensitive to the training set



x

$$\begin{bmatrix} \text{weight} \rightarrow \\ l \longrightarrow \text{mm} \\ b \longrightarrow \text{km} \\ d. \\ \\ \end{bmatrix}$$

large range → d

← large dynamic range dominate

$$\begin{bmatrix} (0,1) \\ \vdots \\ \vdots \end{bmatrix} \rightarrow \text{mean } 0$$

$\sigma^2 = 1$

centered. feature