Lecture 6: Errors - Bias,Variance

27 January 2024 10:58

Recall from last lecture:

$$w^*_{MAP} \in \arg\min_w \left\{ \frac{1}{2\sigma^2}\|Xw-y\|^2 + \frac{\lambda}{2}\|w\|^2 \right\}$$

$$= \frac{1}{\sigma^2}\left(\frac{1}{2\sigma^2}X^TX + \frac{\lambda}{2}I\right)^{-1}X^Ty$$

The matrix $\frac{1}{2\sigma^2}X^TX + \frac{\lambda}{2}I$ is invertible

> A matrix $A_{d\times d}$ is positive definite if, $\forall x \in \mathbb{R}^d \setminus \{0\}$, $x^TAx > 0$
> Any positive definite matrix is invertible

Let $A = \frac{1}{2\sigma^2}X^TX + \frac{\lambda}{2}I$, $v^TAv = \frac{1}{2\sigma^2}v^TX^TXv + \frac{\lambda}{2}\|v\|^2$

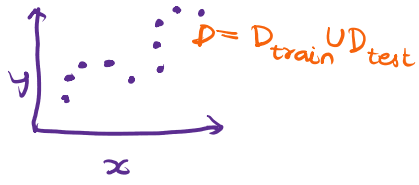
$$= \frac{1}{2\sigma^2}\|Xv\|^2 + \frac{\lambda}{2}\|v\|^2 > 0$$

Prediction Errors:

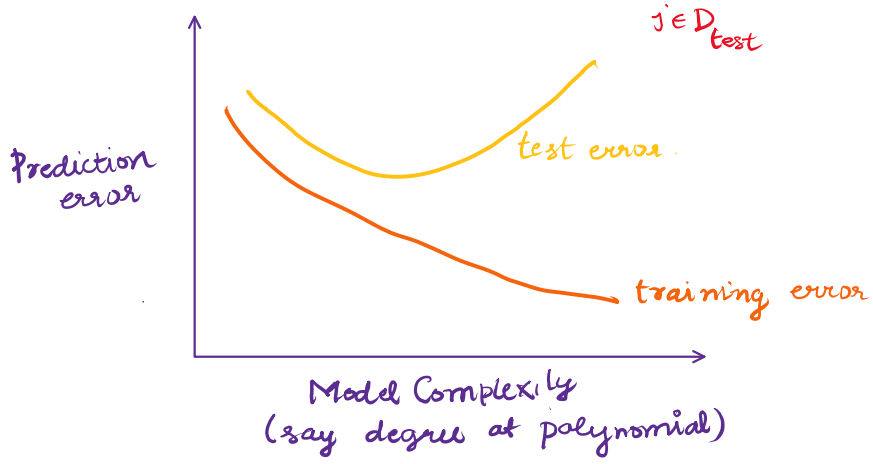
Goal: Estimate $\hat{y}$, which is not present in training example.
we call $(\hat{x}, \hat{y})$ as test data point & $(x_i, y_i) \in D_{train}$ (training data points)
we need our prediction, $g_D(\hat{x})$ to be very close $\hat{y}$.

Measure of goodness:

i) training error: $\sum_{i \in D} \ell(g_D(x_i) - y_i)$ (prediction, actual)



ii) Test error: We hold out a set, D_{test} $\sum_{j \in D_{test}} \ell(g_D(x_i) - y_i)$



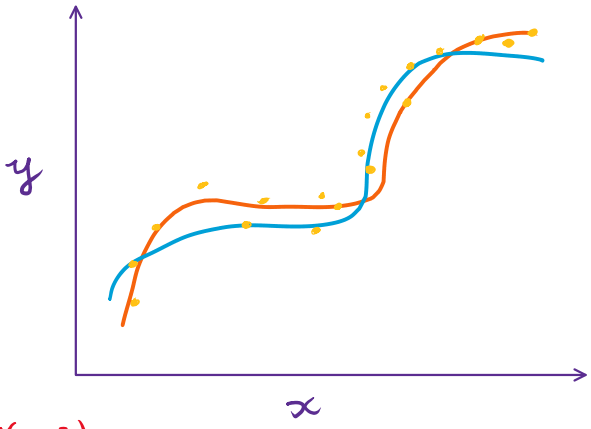
The three primary reasons/causes for test error:

i) Bias ii) Variance iii) Noise (persists always / can't get rid off)

Model dependent

Bias and Variance

Clearly, the predictor function g changes as we change the underlying distribution in which we train it.



$(x_i, y_i) \in D \Rightarrow g_D(x) = \omega^T x$

$(x_i, y_i) \in \tilde{D} \Rightarrow g_{\tilde{D}}(x)$

Let the actual fn be

$$y = f(x) + \epsilon \rightsquigarrow N(0, \sigma^2)$$

We have the test error, $err = g_D(\hat{x}) - \hat{y}$, where $(\hat{x}, \hat{y})$ is a test data point. (assume $\hat{x}$ to be fixed)

NB: 'D' distribution is independent of the $\hat{y}$ distribution.

$$\begin{aligned} err &= g_D(\hat{x}) - \hat{y} \quad \rightarrow A \\ &= \left(g_D(\hat{x}) - \mathbb{E}_D(g_D(\hat{x}))\right) \\ &\quad + \left(\mathbb{E}_D(g_D(\hat{x})) - \mathbb{E}(\hat{y})\right) \rightarrow B = \text{bias} \\ &\quad + \left(\mathbb{E}(\hat{y}) - \hat{y}\right) \rightarrow C \end{aligned}$$

$$\mathbb{E}(err^2) = \mathbb{E}A^2 + \mathbb{E}B^2 + \mathbb{E}C^2 + \underbrace{2\left(\mathbb{E}(AB) + \mathbb{E}(BC) + \mathbb{E}(CA)\right)}_{0}$$

$\mathbb{E}A^2$: Variance, $\mathbb{E}B^2$: Bias^2, $\mathbb{E}C^2$: Noise

NB: $\mathbb{E}(A) = \mathbb{E}(C) = 0$

$\mathbb{E}(B)$ is a constant. (why?)

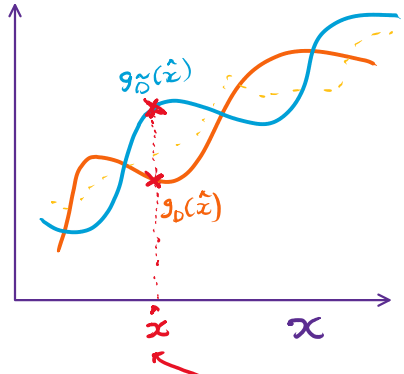
$\mathbb{E}(CA) = \mathbb{E}(C)\,\mathbb{E}(A) = 0$ due to independence

Variance of the model: $\mathbb{E}_D\left[\left(g_D(\hat{x}) - \mathbb{E}(g_D(\hat{x}))\right)^2\right]$

Bias of the model: $\mathbb{E}_D(g_D(\hat{x})) - \mathbb{E}(\hat{y})$

Noise of the model: $\mathbb{E}\left[(\hat{y} - \mathbb{E}(\hat{y}))\right]^2$

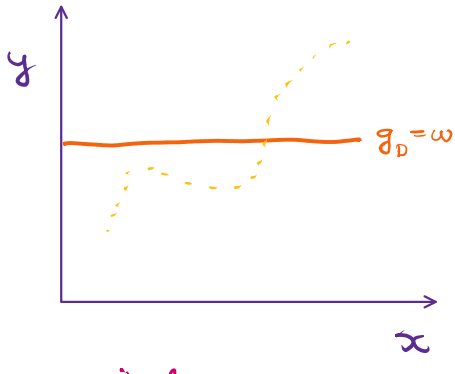
g_D, $g_{\tilde{D}}$: polynomial with degree 10



New point.

i) High variance, low bias

g_D: polynomial with degree 0, $g_D(x) = \omega_0$



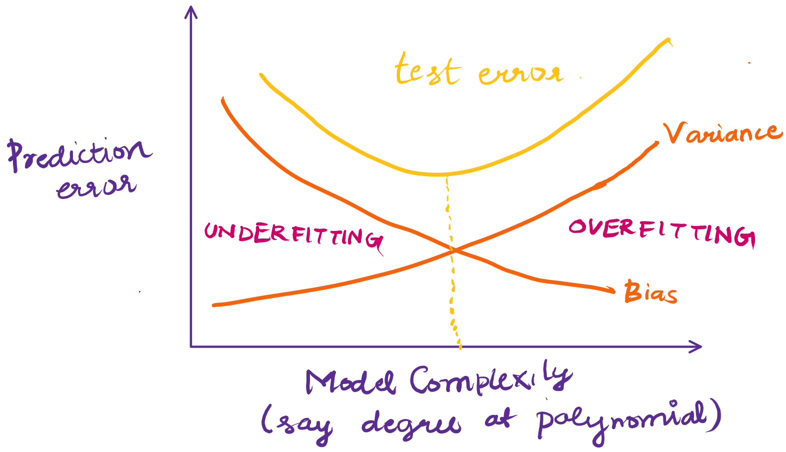
$Var = \mathbb{E}(g_D(\hat{x}) - \mathbb{E}_D(g_D(x))^2$

$= 0$ as g_D is const.

ii) low variance, high bias

(i) High variance, low bias (OVERFITTING) ← New point.

ii) low variance, high bias (UNDERFITTING)



Regularization for linear regression:

Remember our ultimate procedure is optimization:

$$w_{MLE} \in \arg\min \frac{1}{2\sigma^2} \| Xw - y \|^2$$

$$w_{MAP} \in \arg\min \left\{ \frac{1}{2\sigma^2} \| Xw - y \|^2 + \frac{\lambda}{2} \|w\|^2 \right\}$$

We've:

$$\text{Loss}(w) = \text{Loss}_D(w) + \lambda \, \text{Reg}(w)$$

(Regularized model) — λ: hyperparameter; $\text{Reg}(w)$: Regularizer fⁿ

Hyperparameter is usually chosen by trial and error, from a grid of possible values.

Eg: $\text{loss}(w) = E_{\hat{x},\hat{y}} \left(\ell(g_D(\hat{x}), \hat{y}) \right)$

We've following possibilities of Regularizer fⁿ

i) $\text{Reg}(w) = \|w\|_2^2$ (L_2 norm) $= \sqrt{\sum_{i=1}^{d} w_i^2}$

ii) $\text{Reg}(w) = \|w\|_1$ (L_1 norm) $= \sum_{i=1}^{d} |w_i|$

Ridge regression: $$w^* \in \arg\min_{w} \left\{ \| \Phi w - y \|^2 + \lambda \|w\|_2^2 \right\}$$

$$\Phi = \begin{bmatrix} \phi_0(x_1) & \cdots & \phi_m(x_1) \\ \vdots & & \\ \phi_0(x_n) & \cdots & \phi_m(x_n) \end{bmatrix}$$

LASSO regression: $$w^* \in \arg\min_{w} \left\{ \| \Phi w - y \|^2 + \lambda \|w\|_1 \right\}$$

↳ least absolute shrinkage and selection operator

Equivalent optimization problems:

LASSO: $\min_w (\Phi w - y)^T (\Phi w - y)$ s.t. $\|w\|_1 \leq C_1$, ← const depends on λ

Ridge: $\min_w (\Phi w - y)^T (\Phi w - y)$ s.t. $\|w\|_2 \leq C_2$