Chapter 2: Statistical Learning.

- goal: develope an accurate model to predict.

- Observe quantitative response Y
  p different predictors $X = (X_1, X_2, \dots, X_p)$.

Assumption: $Y = f(X) + \varepsilon$

f: fixed but unknown

$\varepsilon \rightarrow$ random ~~error~~ term.

ε: independent of X.

$E(\varepsilon) = 0$.

- f: represents systematic info that X provides about Y.

- 2.1.1 Why estimate f.
  ① Prediction ② Inference.

① **Prediction.**

$\hat{Y} = \hat{f}(X)$. $\hat{f}$: black box.

Accuracy of $\hat{Y}$

$\rightarrow$ Reducible error: improve $\hat{f}$ so that it is close to f, use correct statistical technique.

$\rightarrow$ Irreducible error. $Y = f(X) + \varepsilon$ cause.

even if $\hat{Y} = f(X)$, then ε error.

$$E[Y - \hat{Y}]^2 = E[f(X) + \varepsilon - \hat{f}(X)]^2$$

$$= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{reducible}} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{irreducible}}$$

$E[Y-\hat{Y}]^2$: ~~overall expected~~ variation.

② **Inference.** $\hat{f}$ can not be treated as black box.

1. Which predictors are associated with the response?
↳ identify few important predictors.

2. What is relationship between the response & each predictor?
3. What is complexity of relationship between Y and X_i? is it linear or more complicated?

2.1.2 How Do we Estimate f?

$\hat{f}$ s.t $Y \approx \hat{f}(x)$

Parametric | Non - Parametric

(1) Parametric methods:
1. Assumption about functional form / shape.
$f(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ → linear model.
2. Procedure to fit the model. / train the model.
→ OLS to estimate β_i
"flexible" models cause problem of overfitting?

(2) Non - parametric methods:
- no functional form assumption
- very large no of observations required.
doubt: - "level of smoothness". → may lead to overfitting.

2.1.3 The trade off betn Prediction Accuracy & model interpretability.

- if interested in inference then: restrictive model.
- (relationship with each variable is easy to understand).

2.1.4 Supervised Versus Unsupervised Learning.

Supervised Learning	Unsupervised Learning.
• Response variable present.	• Response Variable absent.
• prediction/ inference.	• Understand relationship betn predictors / find patterns.
• Linear reg, logistic regres, GAM, boosting, etc.	• cluster analysis: whether ~~problems fall into~~ relatively distinct groups. ↳ all classified in correct group.

* Semi-supervised learning: n observ.

m → resp. & pred ~~both available~~

n−m → only predictor available.

~~2.1.5 Regression vs Classification.~~

2.2 Assesing model accuracy:

1. MSE: mean Squared Error.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

→ We choose model with ~~least~~ low test MSE. → one calculated from test data.

- The model with least training MSE may not have low test MSE.

- Fundamental property: model flexibility ↑, training MSE ↓ but test MSE may not.

- overfitting: low train MSE, but high test MSE.

↳ ~~We are finding too much of patterns~~ ~~- capturing patterns due to errors - which are not present in test data.~~

- Cross validation: ~~Method of estimating test MSE using training data.~~

2.2.2 Bias Variance Trade off.

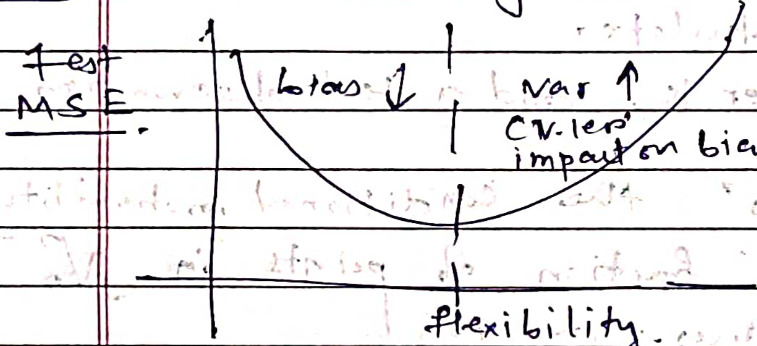
~~E(MSE)~~ =

Expected test MSE = Bias + Var + Var(ε).

$\geq$ Var(ε).

- more flexible statistical methods have higher variance, but less bias.

$\boxed{\text{doubt}}$ How is flexibility measured?



2.2.3 : The classification Setting.

$$\text{error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$ → computes ~~fraction of~~ incorrect classificat^n.

test error rate = Ave $(I(y_0 \neq \hat{y}_0))$.

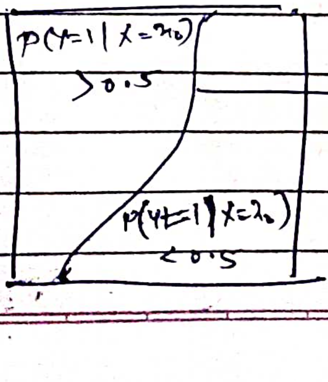
The Bayes Classifier.

assigns each observation to the most likely class, given its predictor values.

$x_0 \to$ class j ~~for~~ if largest $P(Y=j \mid X=x_0)$.

Ex.: $y_i = \{$Class 1, Class 2$\}$.

$x_0 \to$ class 1 if $P(Y=1 \mid X=x_0) > 0.5$.



→ Bayes decision boundary.

$P(Y=1 \mid X=x_0) =$

$P(Y=2 \mid X=x_0) = 0.5$.

- The Bayes classifier produces lowest possible test error rate.
"Bayes error rate".

$$= 1 - E\left[\max_j \{ P(Y=j|X) \}\right]$$

~~We do not know the.~~

K Nearest Neighbors.

- Baye's classifier: we do not know conditional distribution of $Y|X$. → can't calculate.
- Given a positive integer K, and a test observation x_0 represented by N_0.
- It then 'estimates' the conditional probability for class j as the fraction of points in N_0 whose response values equal to j

$$P(Y=j \mid X=x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) = p$$

→ KNN classifies the test observation x_0 to the class with the largest p.

- KNN can often produce classifiers that are surprisingly close to the optimal Bayes classifier.