Post-Midsem L2

Training error in classification

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

↳ $I(y_i \neq \hat{y}) = 0$ (if classified correctly)

$= 1$ (if classified incorrectly)

Test error rate

$$\text{Average}\left(I(y_0 \neq \hat{y}_0)\right)$$

→ Test error can be minimized by a simple Bayes classifier

→ The Bayes classifier assigns each observation to the most likely class, given its predictor value : simply assign a test observation with predictor vector x_0 to the class 'j' for which $P(Y=j \mid X=x_0)$ is maximum over all 'j'

▷ In 2 class senario → 1 $P > 0.5$
→ 2 $P \leq 0.5$

Bayes Error rate = $1 - \underbrace{E\left(\max_j P(y=j|X)\right)}_{\text{Correctness}}$

KNN approximate

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

Bayes Theorem

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

- $P(H|E)$: Prob(Hypotheses such that Evidence is true)
- $P(E|H)$: Prob(Evidence such that Hypotheses is true)
- $P(H)$: Prob(Hypotheses)
- $P(E)$: Prob(Evidence)

$$P(\text{Positive}) = P(D \cap Pos) + P((\sim D) \cap Pos)$$
$$= P(Pos|D) * P(D) + P(Pos|(\sim D)) * P(\sim D)$$

Naive Bayes (or approximate Bayes classifier)

MAP ← Maximum Aposteriori Probability

$$MAP(H) = \max(P(H|E)) = \max\left((P(E|H) * P(H))/P(E)\right)$$

$$P(H|\text{Multiple Evidences}) = \frac{P(E_1|H) * P(E_2|H) * \ldots * P(E_n|H) * P(H)}{P(\text{Multiple Evidences})} \quad -F(P1)$$

Assumption: each of the predictors are independent of each other

Post Midsem-L3

Ex:

<Outlook, temperature, Humidity, Wind>

Predictors

<Play Badminton?> (Yes or No)

$P(No) = \frac{5}{14}$

$P(Yes) = \frac{9}{14}$

(1) $P(\text{Yes}|\text{Sunny}) = \frac{2}{9}$ (Sunny, Yes)

$P(\text{No}|\text{Sunny}) = \frac{3}{5}$ (Sunny, No)

(2)

(2) P(Overcast | Y) = 4/9

P(Overcast | N) = 0/5

(3) P(Rain | Yes) = 3/9

P(Rain | No) = 2/5

(4) P(Hot | Y) = 2/9

P(Hot | N) = 2/5

(5) P(Mild | Y) =

P(Mild | N) =

And so-on

and we put in formula

- F(P1)

Advantages

- Fast, Scalable
- Binary or Multi-class
- Multiple different types
- popular for spam email classification

Disadvantages

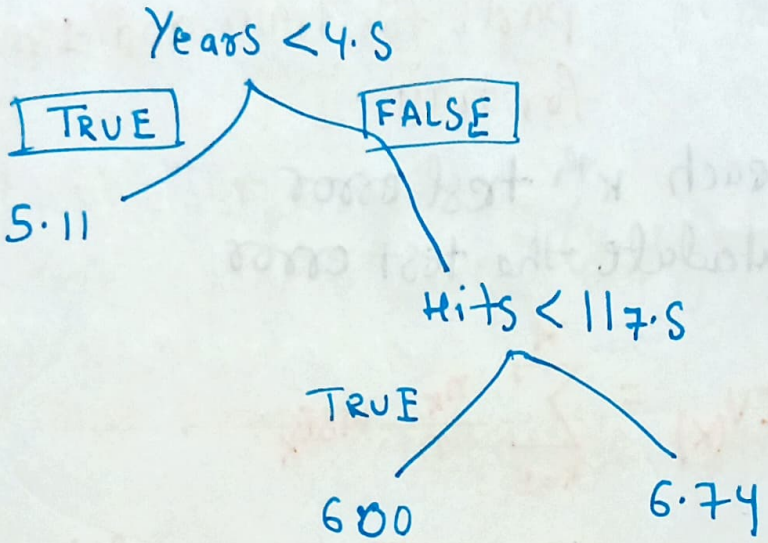
- Assume that predictors are independent.

Tree-based methods

→ Tree based algorithm decides based on multiple division of the attributes (as a tree)

Decision Tree

Ex. [Regression + Classification]



→ Add bagging, random forests and boosting can help increase the accuracy of the decision trees.

Post-Midsem L4

- Estimate test error on basis unavailability of test data
- validation-set (Train/Test) approach has disadvantages.
  - overestimate
  - Not using complete train data

K-Fold Cross-Validation

1, 2 ... K → For each part take 'K-1' parts for train and 1 part for train.

- Using each K^{th} test error we calculate the test error

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$

$$MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$$

{LOOCV} → Leave one out Cross-validation

[Cross-Validation Error]

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2$$

h_i → leverage point

Special case $K = n$ → total number of data points

Taking only one data point in each part

$K = 2$ → Cross validation

So generally $K = 5$ to 10

High leverage → Have a high impact on the model

$$CV_k = \sum_{k=1}^{K} \frac{n_k}{n} E_k$$

$$E_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i) / n_k$$

CV Test Error for logistic regression

Boot Strap

→ Estimate the parameters

Ex:

$$\alpha = \frac{Var(X) - Cov(X,Y)}{Var(X) + Var(Y) - 2\,Cov(X,Y)}$$

Invest in assest return 'X' and $1-\alpha$ in asset returns 'Y'

Post Mid L5

Pruning a tree

→ Larger the decision tree better the results but might possible to overfit the data.

~~A smaller~~

(earlier)

[Some split seems like getting more RSS but might later on could result better RSS.]

T_0 → Prune → Subtree

[Cost Complexity Pruning]

↳ weakest link pruning

T: A subtree of T

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

Tuning parameter

$\hat{y}_{R_m}$ → Predicted value of Region m

$\sum_{i: x_i \in R_m}$ → Decrease with size of T increase

Classification Tree

↳ Discrete values (Logistics)

$$E = 1 - \max_{K}(\hat{p}_{mk})$$

K^{th} class

m^{th} region

Classification error rate ← Not much reliable

Gini index

↳ $G = \sum_{K=1}^{K} \hat{p}_{mK}(1 - \hat{p}_{mK})$

☐ Doubt: Why 'E' is not reliable?