

Home Work - 2

M-Siva Sai Kumar

700773075

1) Decision Sharp

$$h(x) = \begin{cases} + & \text{if sneezing = yes} \\ - & \text{if sneezing = no} \end{cases}$$

Dataset

1. (yes, +) \rightarrow predicted = + \rightarrow correct
2. (no, -) \rightarrow predicted = - \rightarrow correct
3. (yes, -) \rightarrow predicted = + \rightarrow wrong
4. (no, -) \rightarrow predicted = - \rightarrow correct

Step 1: Count

Total samples = 4

Misclassified samples = (read 3 only)

Step 2: Training Error

[error rate] = 25%

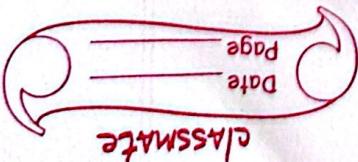
Step 3: compare with memorizer model

If remembers all training data \rightarrow predicts perfectly \rightarrow 0% error

Decision : 25% error

a) Training error = 25%

b) memorizer (0% error)



2.) split on Age (x_1)

young: $\{1, 2\} \rightarrow$ labels: $\{\text{yes}, \text{yes}\}$

majority: yes \rightarrow error in the group = 0

mid : records $\{3, 6\} \rightarrow$ labels: $\{\text{no}, \text{no}\}$

majority = no \rightarrow error = 0

old: $\{4, 5\} \rightarrow$ labels = $\{\text{no}, \text{yes}\}$

* If tie / majority break by majority \rightarrow use

must pick one label majority counts: no, yes \rightarrow tie

* In practice for training - error split, choose the majority label (tie implies any choice causes) with either choice get 1/3 this group

* error = 1

total error = 0 + 0 + 1/3 = 1/3 ≈ 0.333

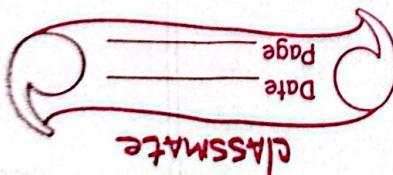
training error = $1/6 \approx 0.1667 \approx 16.67\%$

split on (x_2)

groups:

high: $\{1, 5\} \rightarrow$ labels: $\{\text{yes}, \text{yes}\} \rightarrow$ majority: Yes

medium : records $\{2, 4\} \rightarrow$ labels: $\{\text{yes}, \text{no}\} \rightarrow$ majority: ?



$100 =$ reads $\{3, 6\} \rightarrow$ labels $\{\text{No}, \text{No}\}$ majority

No $\rightarrow \text{error} = 0$

Error $\rightarrow 0 + 1 + 0 = 1$

training error $= 1/6 = 16.67\%$

split on diet ($\times 3$)

Group:

Poor: rec $\{1, 3, 4, 6\} \rightarrow$ labels $= \{\text{Yes}, \text{No}, \text{No}, \text{Yes}\}$

Count: yes=1, No=3 \rightarrow majority = No $\rightarrow \text{error} \neq \text{yes}=1$

Good: rec $\{2, 5\} \rightarrow$ labels $= \{\text{Yes}, \text{Yes}\} \rightarrow$ majority

$\rightarrow \text{Yes} \rightarrow \text{error}=0$

Total error $= 1 + 0 = 1$

1) Training error rates for splitting on each feature

split on Age $\Rightarrow 1/6 = 16.67\%$

split on exercise $\rightarrow 1/6 = 16.67\%$

split on diet $\rightarrow 1/6 = 16.67\%$

2.) All the three features tie with the same splitting order (16.67%) so there is no single best root split by

Page 1

The training error of any of them is equality good under this metric.

(Q3) Entropy Information Gain

Labels: 3 yes, 3 no \rightarrow

$$H(Y) = -[-0.5 \log_2 0.5 + 0.5 \log_2 0.5] = 1.0$$

\rightarrow split on exercise (X_2)

High: (y_{11}, y_{12}) \rightarrow entropy 0

Medium: (y_{21}, y_{22}) \rightarrow entropy 1.0

Low: (y_{31}, y_{32}) \rightarrow entropy 0

$$\text{Weighted Entropy} = (2/6) + 2/6 (1) + 2/6 (0) = 1/3 = 0.333$$

$$\rightarrow \text{Information gain} = 1 - 0.333 = 0.667$$

(Q4) Confusion matrix metrics

Confusion matrix ($T_p = 25, F_n = 5, F_d = 15, T_N = 55, \text{tot} = 100$)

$$\text{Accuracy} = (25+55) / 100 = 0.80$$

$$\text{Precision} = 25 / (25+15) = 0.625$$

$$\text{Recall} = 25 / (25+5) = 0.833$$

$$\text{Specificity} = 55 / (55+15) = 0.786$$

$$F_1 = 2 \cdot (0.625 + 0.833) / (0.625 + 0.33) \approx 0.714$$

If imbalanced (80 negatives 20 positives), because precision are more informative than accuracy.

Q5) distance calculations (kNN)

New point $P(5, 4)$

$$d(P, A) = \sqrt{(5-2)^2 + (4-4)^2} = \sqrt{9} = 3$$

$$d(P, B) = \sqrt{(5-4)^2 + (4-4)^2} = \sqrt{1} = 1$$

$$d(P, C) = \sqrt{(5-4)^2 + (4-6)^2} = \sqrt{5} = 2.236$$

1 NN = nearest neighbour is B \rightarrow predict blue

3 NN = neighbour = {Red, blue, Red} \rightarrow majority &

Q6) k-fold validation

Average error

$$k=1 \rightarrow (0.2 + 0.25 + 0.15 + 0.3) / 4 = 0.225$$

$$k=3 \rightarrow (0.15 + 0.2 + 0.1 + 0.2) / 4 = 0.1625$$

$$k=5 \rightarrow (0.1 + 0.15 + 0.1 + 0.2) / 4 = 0.1375$$

Bgt generalization $k=5$