Homework-2

Lakshmi Prasanna Kamireddy
700772585

Part-A

1) Decision stump

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

Dataset:

1. (Yes, +) → predicted = + → correct
2. (No, -) → predicted = - → correct
3. (Yes, -) → predicted = + → wrong
4. (No, -) → predicted = - → correct

Step-1: Count Errors

total samples = 4

misclassified samples = 1 (record 3 only)

step-2: Training error rate.

$$\text{Error rate} = \frac{\#\text{errors}}{\text{total samples}}$$

$$= \frac{1}{4} = 0.25 = 25\%$$

Step-3:- Compare with memorizer model

Memorizer model: remembers all training data → predicts perfectly → 0% error.

Decision stump: 25% Error

a) Training error rate = 25%.

b) Memorizer is better (0% error).

Q2).

split on Age (x1)

Groups:-

Young : records {1,2} → labels = {yes, yes}

Majority = yes → Errors in this group = 0

Mid : records {3,6} → labels = {No, No}

Majority = No → Errors = 0

Old : records {4,5} → labels = {No, Yes}

* If tie / majority break by majority → we must pick one label:
majority counts : 1 No, 1 Yes → tie.

* In practice for training-error split, choose the majority label (tie implies any choice causes 1 error) with either choice you'll get 1 misclassified in this group.

* Errors = 1

Total Errors = 0 + 0 + 1 = 1

Training Error rate = 1/6 ≈ 0.1667 ≅ 16.67%

Split on Exercise (x2)

Groups:

High : records {1,5} → labels = {yes, yes} → Majority Yes → Errors = 0

Medium : records {2,4} → labels = {yes, No} → Majority ?

tie → whichever label chosen produces 1 error → Error = 1

Low : Records {3,6} → labels = {No, No} → Majority No → Error = 0

Total Errors = 0 + 1 + 0 = 1

Training error rate = 1/6 = 16.67%

split on diet (x3)

Groups :-

Poor : records {1,3,4,6} → Labels = {yes, No, No, No}

Counts : Yes = 1, No = 3 → Majority = No → Errors = # Yes = 1

Good : records {2,5} → Labels = {Yes, Yes} → Majority = yes → Error = 0

Total error = 1 + 0 = 1.

1) Training error rate for splitting on Each feature

split on Age ⇒ 1/6 = 16.67%.

split on Exercise → 1/6 = 16.67%.

split on Diet (x3) → 1/6 = 16.67%.

2) All the three features tie with the same training error (16.67%). So there is no single best root split by the training-error criterion - any of them is equally good under this metric.

Q3) Entropy & Information Gain

→ labels: 3 yes, 3 no →

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

→ Split on Exercise (x2)

High: (Yes, Yes) → Entropy 0

Medium: (Yes, No) → Entropy 1.0

low: (No, No) → Entropy 0

weighted Entropy = $(2/6)0 + (2/6)1 + 2/6(0) = 1/3 = 0.333$

→ Information gain = $1 - 0.333 = 0.667$

Exercise is a good split

Q4) Confusion Matrix metrices:

Confusion matrix ($T_p = 25$, $F_n = 5$, $f_p = 15$, $T_N = 55$, total = 100)

Accuracy = $(25+55)/100 = 0.80$

Precision = $25/(25+55) = 0.625$

Recall = $25/(25+5) = 0.833$

specificity = $55/(55+15) = 0.786$

$F_1 = 2 \cdot (0.625 \times 0.833)/(0625 + 0.33)\ 2 = 0.714$

If imbalanced (80 negatives, 20 positives) Recall, Precision & f_1 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 neighbor is B (Blue) → Predict Blue

3-NN = neighbours = {Red, Blue, Red} → Majority Red

Q6) K-fold Cross-Validation

Average errors:

K=1 → (0.20 + 0.25 + 0.15 + 0.30)/4 = 0.225

K=3 → (0.15 + 0.20 + 0.10 + 0.20)/4 = 0.1625

K=5 → (0.10 + 0.15 + 0.10 + 0.20)/4 = 0.1375

Best Generalization ~~k~~ K=5.