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, +) → predicted = + → correct
2. (no, -) → predicted = - → correct
3. (yes, -) → predicted = + → wrong
4. (no, -) → predicted = - → correct

step1: Count

total samples = 4

Misclassified samples: (record 3 only)

step2: Training Error

[error rate] = 25%

step3: compare with memorizer model

It remembers all training data → predicts perfectly → 0% error

decision : 25% error

a) training error = 25%.

b) memorizer (0% error)

2.) split on Age (x_1)

young: {1,2} → labels {yes, yes}

majority: yes → error in the group: 0

Mid: records {3,6} → labels: {NO, NO}

majority = NO → error: 0

old: {4,5} → labels = {NO, yes}

* If tie / majority break by majority → use must pick on label majority counts: NO, yes → tie

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

* error = 1

total error = 0 + 0 + 1 = 1

Training error = 1/6 ≈ 0.1667 ≈ 16.67%

split on (x_2)

Groups:

High: {1,5} → labels: {yes, yes} → majority: Yes → 0

Medium reads {2, 4} → labels = {yes, no} → majority?

low = reads {3,6} → labels {NO, NO} majority

NO → error=0

Error = 0+1+0=1

training error = 1/6 = 16.67%.

Split on diet (x3)

Group:

poor: rec {1,3,4,6} → labels = {yes, NO, NO, NO}

counts: yes=1, NO=3 → majority = NO → error ≠ yes=1

Good: rec {2,5} → labels = {yes, yes} → majority = yes → error=0

total error = 1+0=1

1) Training error rates for splitting on each feature

split on Age => 1/6 = 16.67%.

split on exercise → 1/6 = 16.67%.

split on diet → 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 any of them is equality good under this metric.

Q3) entropy Information Gain

labels: 3yes, 3no →

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

→ split on exercise (X_2)

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 metrics

Confusion matrix (T_P = 25, F_n = 5, F_D = 15, T_N = 55, Tot = 100)

Accuracy = (25 + 55)/100 = 0.80

precision = 25/(25 + 15) = 0.625

Recall = 25/(25 + 5) = 0.833

Specificity = 55/(55 + 15) = 0.786

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

If imbalanced (80 negatives 20 positives), Recall, 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 → predict blue

3 NN = neighbour = {Red, blue, Red} → majority

Q6) k-fold validation

Average error

$k=1 \rightarrow (0.2 + 0.25 + 0.15 + 0.3)/4 = 0.224$

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

$k=5 \rightarrow (0.1 + 0.15 + 0.1 + 0.2)/4 = 0.137$

Best generalization $k=5$