**University of Central Missouri**  
**Department of Computer Science & Cybersecurity**

**CS5710 Machine Learning**  
**Fall 2025**

**Home Assignment 2 – Part A Solutions**

**Student Name: Gangadhar Nersu(700765662)**

**Part A: Calculation (Questions 1–6)**

**Q1. Decision Stump Prediction**

**Decision stump:** h(x) = + if Sneezing = Yes, - otherwise.

**Dataset:**

1. (Sneezing=Yes, Label=+)
2. (Sneezing=No, Label=-)
3. (Sneezing=Yes, Label=-)
4. (Sneezing=No, Label=-)

**Step 1: Predictions**

* Record 1: Yes → + (Correct)
* Record 2: No → - (Correct)
* Record 3: Yes → + (Incorrect)
* Record 4: No → - (Correct)

**Step 2: Errors**

* Total records: 4
* Errors: 1 (record 3)

**Training error rate:** 1/4 = 0.25 (25%)

**Memorizer model:** predicts perfectly → error = 0%

**Comparison:** The decision stump has 25% training error, whereas the memorizer has 0%.

**Q2. Training Error as Splitting Criterion**

**Dataset:**

| **Age (x1)** | **Exercise (x2)** | **Diet (x3)** | **Label** |
| --- | --- | --- | --- |
| Young | High | Poor | Yes |
| Young | Medium | Good | Yes |
| Mid | Low | Poor | No |
| Old | Medium | Poor | No |
| Old | High | Good | Yes |
| Mid | Low | Poor | No |

**Step 1: Training error for each feature split**

**Split on Age (x1)**

* Young: 2 records (Yes, Yes) → predict Yes → 0 errors
* Mid: 2 records (No, No) → predict No → 0 errors
* Old: 2 records (No, Yes) → predict majority No → 1 error
* Total errors = 1/6 → 16.7%

**Split on Exercise (x2)**

* High: 2 records (Yes, Yes) → predict Yes → 0 errors
* Medium: 2 records (Yes, No) → predict majority Yes → 1 error
* Low: 2 records (No, No) → predict No → 0 errors
* Total errors = 1/6 → 16.7%

**Split on Diet (x3)**

* Poor: 4 records (Yes, No, No, No) → predict majority No → 1 error
* Good: 2 records (Yes, Yes) → predict Yes → 0 errors
* Total errors = 1/6 → 16.7%

**Step 2: Best feature:** All three features give equal training error (16.7%).

**Q3. Entropy & Information Gain**

**Step 1: Entropy of labels**

* Labels: Yes = 3, No = 3
* Entropy formula: H=−∑pilog⁡2piH = -\sum p\_i \log\_2 p\_i
* pYes=3/6=0.5, pNo= 0.5
* H(Label)=−0.5∗log2(0.5)−0.5∗log2(0.5)=1.0

**Step 2: Entropy after splitting on Exercise (x2)**

* High: Yes=2, No=0 → H=0
* Medium: Yes=1, No=1 → H=-0.5log2(0.5)-0.5log2(0.5)=1
* Low: Yes=0, No=2 → H=0
* Weighted entropy: (2/6)∗0+(2/6)∗1+(2/6)∗0=1/3≈0.333

**Step 3: Information Gain**

* IG = H(Label) - H(Label|Exercise) = 1 - 0.333 = 0.667

**Step 4: Is Exercise a good split?**

* IG = 0.667 (high) → Yes, it is a good split.

**Q4. Confusion Matrix Metrics**

**Confusion matrix:**

|  | **Pred +** | **Pred -** |
| --- | --- | --- |
| Actual + | 25 | 5 |
| Actual - | 15 | 55 |

**Step 1: Metrics**

* Accuracy = (TP+TN)/Total = (25+55)/100 = 0.8
* Precision = TP/(TP+FP) = 25/(25+15) = 0.625
* Recall = TP/(TP+FN) = 25/(25+5) = 0.833
* Specificity = TN/(TN+FP) = 55/(55+15) = 0.785
* F1-score = 2\*(Precision*Recall)/(Precision+Recall) = 2*0.625\*0.833/(0.625+0.833) ≈ 0.714

**Step 2: Imbalanced dataset**

* If negatives dominate (80 neg, 20 pos), Accuracy can be misleading.
* **Most informative metric:** F1-score (balances precision & recall for minority class)

**Q5. Distance Calculations (kNN)**

**Points:**

* A(2,4), Red
* B(4,4), Blue
* C(4,6), Red
* New point P(5,4)

**Step 1: Euclidean distances**

* d(P,A) = sqrt((5-2)^2 + (4-4)^2) = sqrt(9 + 0) = 3
* d(P,B) = sqrt((5-4)^2 + (4-4)^2) = sqrt(1 + 0) = 1
* d(P,C) = sqrt((5-4)^2 + (4-6)^2) = sqrt(1 + 4) = sqrt(5) ≈ 2.236

**Step 2: 1-NN prediction**

* Nearest neighbor: B (Blue) → P = Blue

**Step 3: 3-NN prediction**

* 3 nearest: B(Blue), C(Red), A(Red)
* Majority: Red → P = Red

**Q6. K-fold Cross-Validation**

**Errors:**

| **Fold** | **k=1** | **k=3** | **k=5** |
| --- | --- | --- | --- |
| 1 | 0.20 | 0.15 | 0.10 |
| 2 | 0.25 | 0.20 | 0.15 |
| 3 | 0.15 | 0.10 | 0.10 |
| 4 | 0.30 | 0.20 | 0.20 |

**Step 1: Mean CV error**

* 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

**Step 2: Best generalization**

* Lowest mean CV error = k=5 → k=5 generalizes best.