CS 6350/DS 4350: HW1 – Q2 – Decision Trees on the University (Scholarship) Dataset

Submission date: 9/9/2025

Group 2

Name: Jackson Switzer

Student ID: u1608753

**Grading Rubric**

| **Question** | **Points** | **Scoring rules** | **Graded Points** |
| --- | --- | --- | --- |
| 1-(a) | 8 | Total Boolean functions – 2 pts  Reasoning for total functions – 2 pts Consistent functions – 2 pts  Reasoning for consistent functions – 2 pts |  |
| 1-(b) | 6 | Final entropy result – 2 pts  Reasoning process – 4 pts |  |
| 1-(c) | 10 | Every feature result – 1 pt (total 4 pts)  Reasoning process of GPA/Extracurricular – 4 pts Reasoning process of Recommendation/Year – 2 pts |  |
| 1-(d) | 4 | The root result – 2 pts  Reasoning process – 2 pts |  |
| 1-(e) | 12 | A textual description or a simple diagram – 6 pts  Reasoning process – 6 pts |  |
| 1-(f) | 2 | Result – 1 pt  Reasoning process – 1 pt |  |
| 1-(g) | 8 | Every feature result (0.5 pts) and reasoning process (1pt) – 1.5 pts each (total 6 pts)  The root result – 1 pt  Compare with entropy and Gini – 1 pt |  |
| **Total:** | **50** |  |  |

**Problem 1**

A university wants to decide whether to offer a scholarship to an applicant. The decision depends on the following four features, analogous to classic decision-tree exercises (e.g., the weather example):

* GPA (Low, Medium, High)
* Extracurricular (None, Some, Strong)
* Recommendation (Weak, Strong)
* Year (Junior, Senior)

You are given **10 past applications** (Table 1). For each application, the four feature values and the final scholarship decision are recorded.

Table 1: Training data for the scholarship prediction problem.

| GPA | Extracurricular | Recommendation | Year | Scholarship |
| --- | --- | --- | --- | --- |
| High | Strong | Strong | Senior | + |
| Low | None | Weak | Junior | - |
| Medium | Some | Strong | Senior | + |
| Low | Strong | Weak | Junior | - |
| High | Some | Weak | Senior | + |
| Medium | None | Weak | Senior | - |
| High | None | Strong | Junior | + |
| Medium | Strong | Strong | Junior | + |
| Low | Some | Strong | Senior | - |
| Medium | Strong | Weak | Senior | + |

1. [8 pts] How many Boolean functions are possible over the input space? How many are consistent with the given dataset (10 labeled rows)? Show your reasoning.

**Solution:**

There are 3 possible values of GPA, 3 possible values of Extracurricular, 2 possible values of Recommendation, and 2 possible values of Year, resulting in 3\*3\*2\*2=36 possible feature combinations in the input space.

**The number of boolean functions possible over that input space is 2number of instances**

**= 236 = 68,719,476,736**

With the given dataset, 10 rows are fixed, so only the 26 remaining rows can vary, meaning **the number of boolean functions that are possible over the input space *and* are consistent with the given data is 226 = 67,108,864**

1. [6 pts] Compute the entropy of the labels in Table 1. Show your counts and the formula.

**Solution:**

With P(i) denoting the probability of label i occurring in the dataset:

With the given data, P(-) = 4/10 = 0.4 and P(+) = 6/10 = 0.6, hence

1. [10 pts] Compute the **information gain (using entropy)** for each feature (GPA, Extracurricular, Recommendation, Year). Report the results in Table 2.

**Solution:**

H(X where GPA=low) = 0 since all instances have the same label (-)

H(X where GPA=medium) = -[3/4\*log2(3/4) + 1/4\*log2(1/4)] ≈ 0.81128

H(X where GPA=high) = 0 since all instances have the same label (+)

H(X where extracurricular=none) = -[2/3\*log2(2/3) + 1/3\*log2(1/3)] ≈ 0.91830

H(X where extracurricular=some) = -[1/3\*log2(1/3) + 2/3\*log2(2/3)] ≈ 0.91830

H(X where extracurricular=strong) = -[3/4\*log2(3/4) + 1/4\*log2(1/4)] ≈ 0.81128

H(X where recommendation=weak) = -[3/5\*log2(3/5) + 2/5\*log2(2/5)] ≈ 0.97095

H(X where recommendation=strong) = -[4/5\*log2(4/5) + 1/5\*log2(1/5)] ≈ 0.72193

H(X where year=junior) = -[1/2\*log2(1/2) + 1/2\*log2(1/2)] = 1

H(X where year=senior) = -[2/3\*log2(2/3) + 1/3\*log2(1/3)] ≈ 0.91830

IG(X split on GPA) = 0.97095 - [(3/10)(0) + (4/10)(0.81128) + (3/10)(0)] = **0.64644**

IG(X split on extracurricular) = 0.97095 - [(3/10)(0.91830) + (3/10)(0.91830) + (4/10)(0.81128)]

= **0.09546**

IG(X split on recommendation) = 0.97095 - [(1/2)(0.97095) + (1/2)(0.72193)] = **0.12451**

IG(X split on year) = 0.97095 - [(2/5)(1) + (3/5)(0.91830)] = **0.01997**

Graphic 11

| Table 2: Information gain (entropy) for each feature. | |
| --- | --- |
| Feature | Information Gain |
| GPA | 0.646 |
| Extracurricular | 0.095 |
| Recommendation | 0.125 |
| Year | 0.020 |

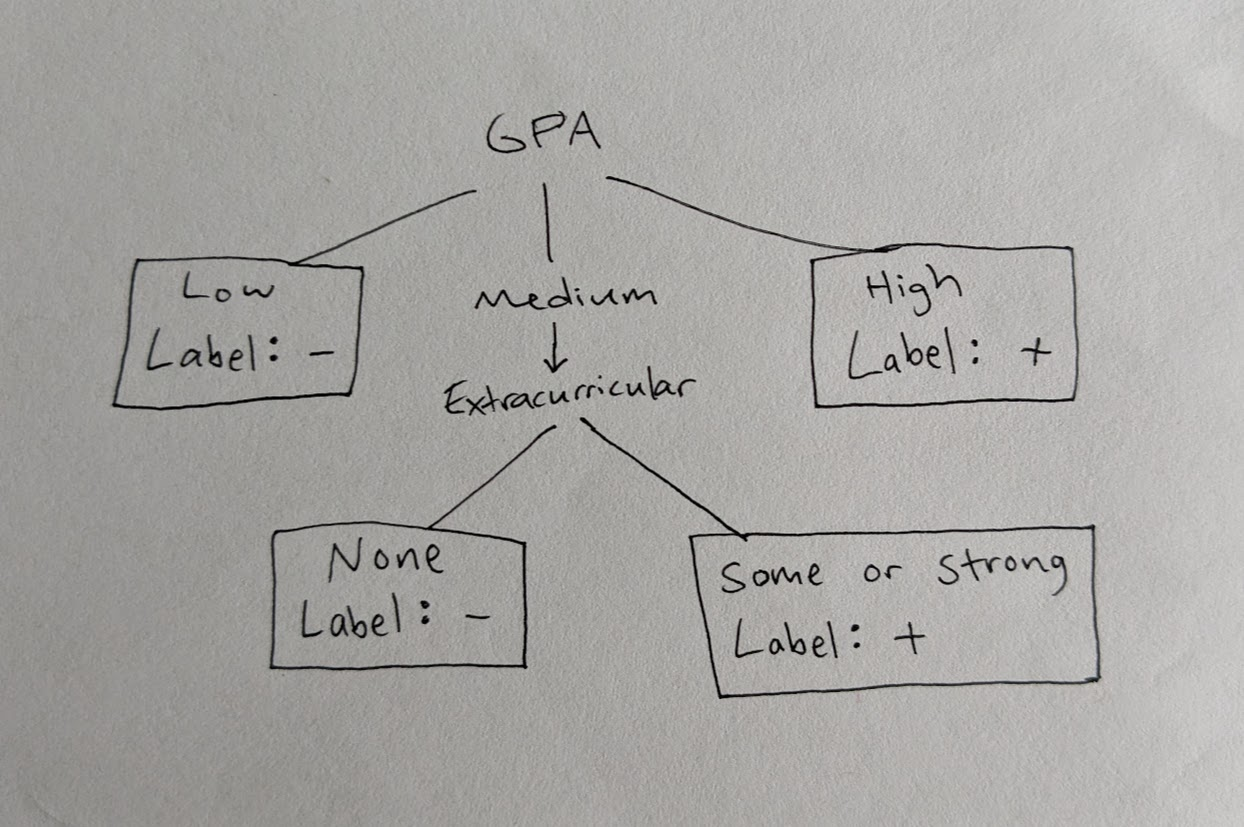
1. [4 pts] Based on Table 2, which attribute would **ID3** choose as the root? Justify briefly.

**Solution**:

ID3 would choose GPA as the root because it is a greedy algorithm, so it makes the locally optimal choice. The tree gains the most information on the first split if it splits on GPA, so that’s what it would do.

1. [12 pts] Construct a decision tree using your chosen root that correctly classifies the training data. A textual description or a simple diagram is acceptable.

**Solution:**

****

I did it this way because GPA alone categorizes most of the data, so it seemed like the best first feature to split on. Of the data points with Medium GPA, only one was labeled -, so I found one of the features that distinguished that point from the others: that point was the only one with Medium GPA and “None” for extracurriculars. So Extracurricular/None became a leaf with label -, and the other two labels for extracurricular I combined to form a leaf with label +.

There were other options, such as separating Extracurricular/Some and Extracurricular/Strong into their own leafs, or using Recommendation as the second splitting feature instead of Extracurricular, but this seemed like the simplest option, so it’s what I did. It might underfit the data, but it’s hard to tell without more instances.

1. [2 pts] Test your tree on the **3 new applications** in Table 3 and report the accuracy.

Table 3: Test data for the scholarship prediction problem.

| GPA | Extracurricular | Recommendation | Year | Scholarship |
| --- | --- | --- | --- | --- |
| High | None | Weak | Senior | + |
| Low | Strong | Strong | Junior | - |
| Medium | Some | Weak | Senior | - |

**Solution:**

First row: GPA High => Label + correct

Second row: GPA Low => Label - correct

Third row: GPA Medium, Extracurricular Some => Label - incorrect

Accuracy = 66.66%

1. [8 pts] **Gini impurity as a split criterion (within the 50 points).**

Gini(Z) = 1 − Σ p\_i^2, where p\_i is the fraction of class i in set Z.

Define:

Gain\_Gini(A) = Gini(Y) − Σ\_v P(A=v) · Gini(Y | A=v)

Compute Gain\_Gini for each feature and state which attribute Gini would choose as the root. Do

entropy and Gini agree?

**Solution:**

Gini(Y) =

Gini(Y where GPA=low) = 1 - (12 + 02 + 02) = 0

Gini(Y where GPA=medium) = 1 - [(3/4)2 + (1/4)2] = 0.375

Gini(Y where GPA=high) = 1 - [02 + 02 + 12] = 0

Gini(Y where extracurricular=none) = 1 - [(2/3)2 + (1/3)2] = 0.444

Gini(Y where extracurricular=some) = 1 - [(1/3)2 + (2/3)2] = 0.444

Gini(Y where extracurricular=strong) = 1 - [(3/4)2 + (1/4)2] = 0.375

Gini(Y where recommendation=weak) = 1 - [(3/5)2 + (2/5)2] = 0.48

Gini(Y where recommendation=strong) = 1 - [(4/5)2 + (1/5)2] = 0.32

Gini(Y where year=junior) = 1 - [(1/2)2 + (1/2)2] = 0.5

Gini(Y where year=senior) = 1 - [(2/3)2 + (1/3)2] = 0.444

Gain\_Gini(Y split on GPA) = 0.48 - [(3/10)(0) + (4/10)(0.375) + (3/10)(0)] = 0.33

Gain\_Gini(Y split on extracurricular) = 0.48 - [(3/10)(0.444) + (3/10)(0.444) + (4/10)(0.375)] = 0.064

Gain\_Gini(Y split on recommendation) = 0.48 - [(1/2)(0.48) + (1/2)(0.32)] = 0.08

Gain\_Gini(Y split on year) = 0.48 - [(4/10)(0.5) + (6/10)(0.444)] = 0.014

Gini would choose GPA as the root; it agrees with entropy.