Homework 2

Automated Learning and Data Analysis
Dr. Thomas Price

Spring 2021

Instructions

Due Date: March, 5 2021 at 11:45 PM
Total Points: 60 for CSC522 and CSC422
Submission checklist:

- Clearly list each team member's names and Unity IDs at the top of your submission.
- Your submission should be a single PDF file containing your answers. **Name your file**: G(homework group number)_HW(homework number), e.g. G1_HW2.
- If a question asks you to explain or justify your answer, **give a brief explanation** using your own ideas, not a reference to the textbook or an online source.
- Submit your PDF through Gradescope under the HW2 assignment (see instructions on Moodle). **Note**: Make sure to add you group members at the end of the upload process.
- In addition to your group submission, please also *individually* submit your Programming portion via our JupyterHub site *and* Moodle.

1 Decision Tree Construction (20 points) [Chengyuan]

Create decision trees **by hand** for the `hw2q1.csv` Titanic survival dataset, as explained below, using Hunt's algorithm. Note the following:

- In the given dataset, all of the input attributes are binary except for the "Pclass" which is categorical. "Pclass" should create a 3-way split if used in the tree.
- The output label has two class values: T or F, which represent Survival or Not Survival.
- In the case of ties when selecting an attribute, break ties in favor of the leftmost attribute.
- When considering a split for the continuous attribute, identify the best value to split on (e.g. ≤ 15 and > 15) by testing all possible split values.

You must show your work when calculating Information Gain or Gini Index for *the split at the root node* (but not for later splits). You can do so by either 1) writing out subseps (e.g. conditional entropy for each child node), or including a code for a program you used to make your calculations.

You should draw a separate tree, like the example in Figure 1, after each attribute split. You can use a program (e.g. tikz with LaTeX, Lucidchart, etc.) to draw your trees, or draw them by hand on paper and scan your results into the final pdf.

1a. Construct the decision tree manually, using Gini index to select the best attribute to split on. The maximum depth of your tree should be 3 (count the root node as depth 0), meaning that any node at depth 3 will automatically be a leaf node, even if it has objects with different classes.

1b. Construct the tree manually using Information Gain. The maximum depth of the tree should be 2.

1c. How are the trees different? As part of your explanation, give 1 examples of data objects that would be classified differently by the two trees.

1d. Which decision tree will perform better on the training dataset (`hw2q1.csv`)? Which will perform better on a test dataset? Can we know the answer?

2 Evaluation Measures & Pruning (13 points) [Chengyuan]

This analysis pertains to the *IBM Attrition* dataset, which includes attributes about employess and whether they left the company (Yes/No). The main goal of the analysis is to study the indicators of attrition in order to identify ways that the company can improve employee retention to save money and time spent in hiring and training. To predict the attrition, consider using the decision tree shown in Figure 1 which involves Business Travel Frequency (BTF), Gender, Marital Status (MS) and Engineer Level (EL). Complete the following tasks:

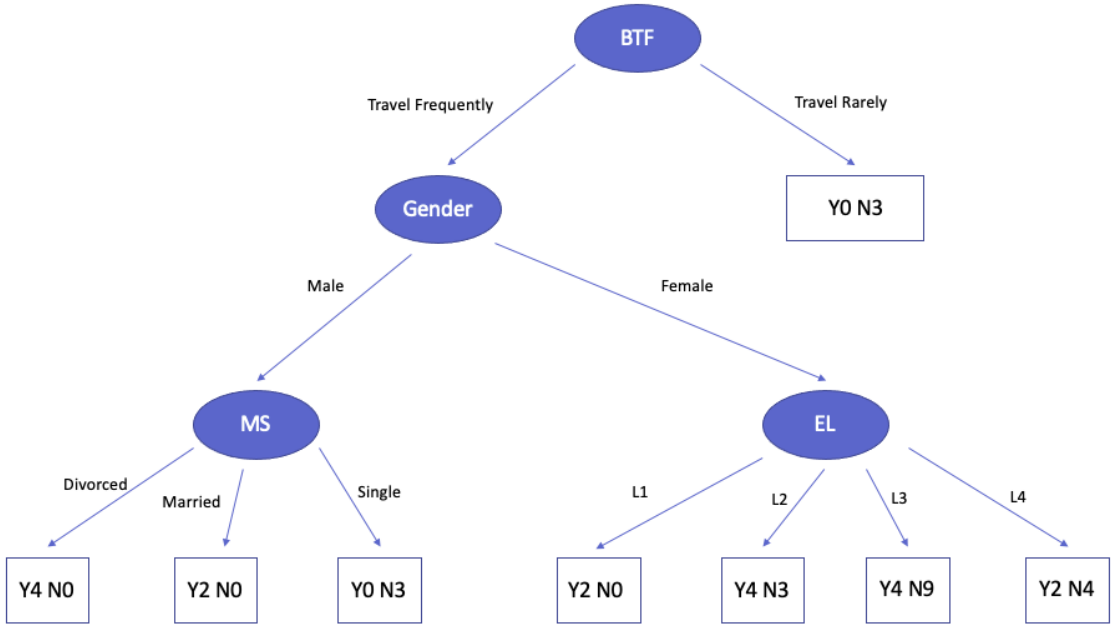

Figure 1: Decision Tree

2a. Use the decision tree above to classify the provided dataset. `hw2q2.csv`. Construct a confusion matrix and report the test Accuracy, Error Rate, Precision, Recall, and F1 score. Use "Yes" as the positive class in the confusion matrix.

2b. Calculate the optimistic training classification error before splitting and after splitting using **EL**, respectively. **Consider only the subtree starting with the EL node.** If we want to minimize the optimistic error rate, should the node's children be pruned?

2c. Calculate the pessimistic training errors before splitting and after splitting using **EL** respectively. Consider only the subtree starting with the EL node. When calculating pessimistic error, use a leaf node error penalty of 0.8. If we want to minimize the pessimistic error rate, should the node's children be pruned?

2d. Assuming that the "EL" node is pruned, recalculate the test Error Rate using `hw2q2.csv`. Based on your evaluation using the test dataset in `hw2q2.csv`, was the original tree (with the EL node) over-fitting? Why or why not?

3 1-NN, & Cross Validation (15 points) [Krishna Gadiraju]

Consider the following dataset (9 instances) with **2 continuous attributes** (x_1 and x_2) that have been scaled to be in the same range, and a **class attribute** y, shown in Table 1. For this question, we will consider a 1-Nearest-Neighbor (1-NN) classifier that uses euclidean distance.

Table 1: 1-NN

ID	x1	x2	Class
1	5.56	1.25	+
2	3.61	3.33	-
3	8.06	5	-
4	3.89	4.17	+
5	10	7.5	-
6	2.78	7.08	-
7	1.94	0	+
8	2.22	6.25	+
9	6.11	4.17	-

3a. Calculate the distance matrix for the dataset using euclidean distance. **Tip**: You can write a simple program to do this for you, and there is an example of how to do this in the programming portion of this homework.

3b. By hand, evaluate the 1-NN classifier, calculating the confusion matrix and testing accuracy (show your work by labeling each data object with the predicted class). **Tip**: you can scan a row or column of the distance matrix to easily find the closest neighbor. Use the following evaluation methods:

i) A holdout test dataset consisting of last 4 instances

ii) 3-fold cross-validation, using the following folds with IDs: [1,2,3], [4,5,6], [7,8,9] respectively.

iii) Leave one out cross validation (LOOCV)

3c. For a data analysis homework, you are asked to perform an experiment with a binary classification algorithm. You are given a dataset with 50 instances and a class attribute that can be either Positive or Negative. The dataset includes 25 positive and 25 negative instances. You use three different validation methods: holdout (with a random 30/20 training/validation split), 5-fold cross validation (with random folds) and LOOCV. As a baseline, you compare your algorithm to a "simple majority classifier," which always predicts the majority class in the training dataset (if there is no majority, one of the classes is chosen at random). You expect the simple majority classifier to achieve approximately 50% validation accuracy, but for one of these evaluation methods you get 0% validation accuracy. Which evaluation gives this results and why?

4 BN Inference (12 points) [Krishna Gadiraju]

The following dataset presents 3 categorical attributes: Gender (M, F), Car Type (Sports, Luxury) and Age Group (G1, G2) with one Class Variable: Class (C0, C1). For each question, please show how you arrived at your answer.

Gender	**Car Type**	**Age Group**	**Class**
M	Luxury	G2	C0
M	Sports	G1	C0
M	Sports	G1	C1
M	Luxury	G1	C1
M	Luxury	G2	C0
F	Sports	G1	C1
F	Luxury	G2	C1
F	Luxury	G1	C0
F	Sports	G1	C0
F	Luxury	G1	C1

Table 2: Dataset for BN Inference

For the following problem, you may find it useful to fill in the following table (optional).

$P(Class = C0) =$	$P(Class = C1) =$
$P(Gender = M \mid Class = C0) =$	$P(Gender = M \mid Class = C1) =$
$P(Gender = F \mid Class = C0) =$	$P(Gender = F \mid Class = C1) =$
$P(CarType = Luxury \mid Class = C0) =$	$P(CarType = Luxury \mid Class = C1) =$
$P(CarType = Sports \mid Class = C0) =$	$P(CarType = Sports \mid Class = C1) =$
$P(AgeGroup = G1 \mid Class = C0) =$	$P(AgeGroup = G1 \mid Class = C1) =$
$P(AgeGroup = G2 \mid Class = C0) =$	$P(AgeGroup = G2 \mid Class = C1) =$

Using the training dataset above, how would a Naive Bayes classifier classify the following data points? Show your work.

4a. $\{Gender = M,\ Car\ Type = Luxury,\ Age\ Group = G1\}$

4b. $\{Gender = M,\ Car\ Type = Sports,\ Age\ Group = G2\}$

4c. $\{Gender = F,\ Car\ Type = Sports,\ Age\ Group = G1\}$

4d. $\{Gender = F,\ Car\ Type = Luxury,\ Age\ Group = G2\}$