

# Homework 2

Automated Learning and Data Analysis  
Dr. Thomas Price

Spring 2022

## Instructions

**Due Date:** February, 28 2022 at 11:45 PM

**Total Points:** 40

**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.
- Submit the programming portion of the homework *individually* through JupyterHub.

# 1 Evaluation Measures & Pruning (13 points) [Chengyuan]

This analysis pertains to the *IBM Attrition* dataset, which includes attributes about employees 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:

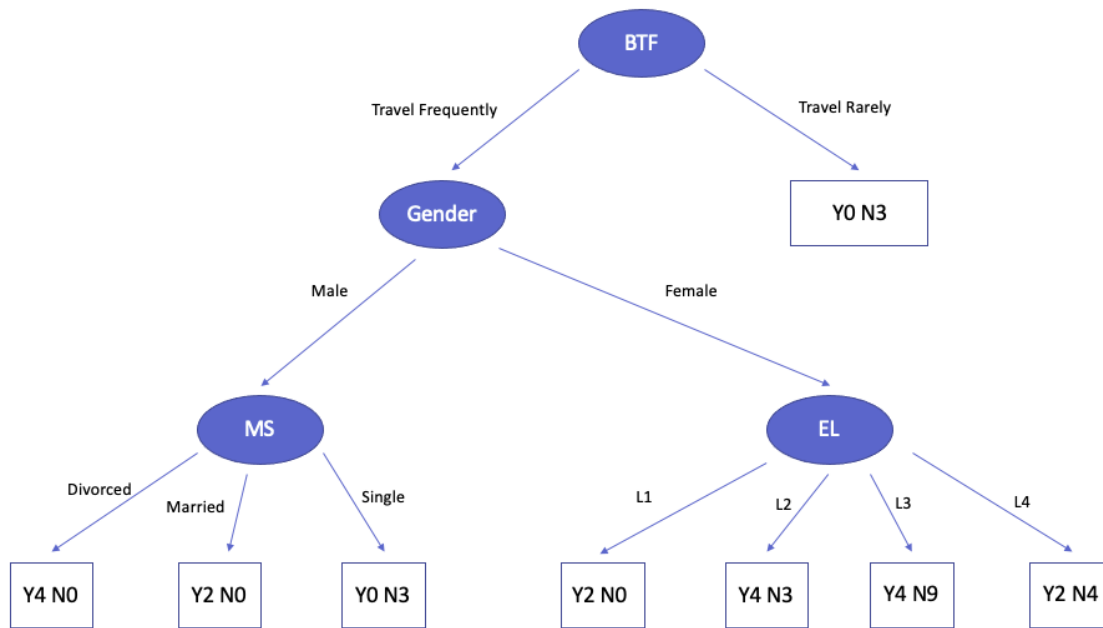


Figure 1: Decision Tree

- 1a) (3 points) Use the decision tree above to classify the provided dataset. `hw2q1.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.
- 1b) (4 points) 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?
- 1c) (4 points) 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?
- 1d) (2 points) 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?

## 2 1-NN, & Cross Validation (15 points) [Jianxun WANG]

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	4.23	7.01	-
2	2.15	0.12	+
3	5.33	5.14	-
4	3.49	9.64	+
5	0.15	4.23	-
6	8.23	1.22	+
7	6.48	2.09	-
8	1.53	2.32	+
9	4.78	3.67	-

- 2a) (3 points) Calculate the distance matrix for the dataset using euclidean distance. **Tip:** You can write a simple program to do this for you.
- 2b) (9 points) 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).
- 2c) (3 points) For a data analysis homework, you are asked to perform an experiment with a binary classification algorithm that uses a “simple majority vote 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 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. 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?

### 3 BN Inference (12 points) [Benyamin Tabarsi]

The following dataset presents 3 categorical attributes: Color (Black, White), Car Type (Sports, Luxury) and Body Style (Sedan, SUV) with one class Variable: Popular (Yes, No). For each question, please show how you arrived at your answer.

Color	Car Type	Body Style	Popular
Black	Luxury	Sedan	Yes
Black	Sports	Sedan	No
Black	Sports	SUV	Yes
White	Luxury	Sedan	Yes
White	Luxury	SUV	No
White	Luxury	Sedan	No
Black	Luxury	SUV	Yes
Black	Luxury	Sedan	No
White	Sports	SUV	No
Black	Luxury	SUV	Yes

Table 2: Dataset for BN Inference

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

$P(\text{Popular} = \text{Yes}) =$	$P(\text{Popular} = \text{No}) =$
$P(\text{Color} = \text{Black} \mid \text{Popular} = \text{Yes}) =$	$P(\text{Color} = \text{Black} \mid \text{Popular} = \text{No}) =$
$P(\text{Color} = \text{White} \mid \text{Popular} = \text{Yes}) =$	$P(\text{Color} = \text{White} \mid \text{Popular} = \text{No}) =$
$P(\text{CarType} = \text{Luxury} \mid \text{Popular} = \text{Yes}) =$	$P(\text{CarType} = \text{Luxury} \mid \text{Popular} = \text{No}) =$
$P(\text{CarType} = \text{Sports} \mid \text{Popular} = \text{Yes}) =$	$P(\text{CarType} = \text{Sports} \mid \text{Popular} = \text{No}) =$
$P(\text{BodyStyle} = \text{Sedan} \mid \text{Popular} = \text{Yes}) =$	$P(\text{BodyStyle} = \text{Sedan} \mid \text{Popular} = \text{No}) =$
$P(\text{BodyStyle} = \text{SUV} \mid \text{Popular} = \text{Yes}) =$	$P(\text{BodyStyle} = \text{SUV} \mid \text{Popular} = \text{No}) =$

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

- 3a) (3 points)  $\{\text{Color} = \text{Black}, \text{Car Type} = \text{Luxury}, \text{Body Style} = \text{Sedan}\}$
- 3b) (3 points)  $\{\text{Color} = \text{Black}, \text{Car Type} = \text{Sports}, \text{Body Style} = \text{SUV}\}$
- 3c) (3 points)  $\{\text{Color} = \text{White}, \text{Car Type} = \text{Sports}, \text{Body Style} = \text{Sedan}\}$
- 3d) (3 points)  $\{\text{Color} = \text{White}, \text{Car Type} = \text{Luxury}, \text{Body Style} = \text{SUV}\}$