

## CS 486/686: Assignment 2 (100 Points)

Instructors: Jesse Hoey & Victor Zhong

Due Date: Wednesday February 26th, 2025 at 11:59 pm (ET; Waterloo Time)

### Instructions

- Submit written solutions in a file named `writeup.pdf` and code in a file named `code.py` to the A2 Dropbox on Learn (<https://learn.uwaterloo.ca>). **If your written submission on Learn does not contain `one` file named `writeup.pdf`, and one file named `code.py` we will deduct 5 marks from your final assignment mark.**
- No late assignment will be accepted.
- This assignment is to be done individually.
- Lead TAs:
  - Varun Dhanraj (vdhanraj)
  - Tien Dat Nguyen (tdnguyen)

The TA office hours will be scheduled on Piazza.

## 1 Question 1 [60pts]: Text Categorization with Decision Trees

Text categorization is an important task in natural language processing and information retrieval. For instance, news articles, emails or blogs are often classified by topics. In this assignment, you will implement (in the language of your choice) a decision tree algorithm to learn a classifier that can assign a discussion board topic to an article.

Train and test your algorithm with the provided datasets in `trainData.txt` and `testData.txt`. The dataset includes the first 1500 comments of August 2019 of each of the `r/books` and `r/atheism` subreddits, cleaned by removing punctuation and some offensive language, and limiting the words to only those used more than 3 times among all posts. These 3000 comments are split evenly into training and testing sets (with 1500 documents in each). To simplify your implementation, these posts have been pre-processed and converted to the *bag of words* model. More precisely, each post is converted to a vector of binary values such that each entry indicates whether the document contains a specific word or not.

Each line of the files `trainData.txt` and `testData.txt` are formatted as “`docId wordId`”, which indicates that word `wordId` is present in document `docId`. Files `trainLabel.txt` and `testLabel.txt` indicate the label/category (1=atheism or 2=books) for each document (`docId` = line#). The file `words.txt` indicates which word corresponds to each `wordId` (denoted by the line#).

Your code will need to above information from the provided files.

Implement a decision tree learning algorithm. Here, each decision node corresponds to a word feature, and the leaf nodes correspond to a prediction of what subreddit the article belongs in. You will experiment with two feature selection mechanisms as follows:

1. Average information gain (evenly weighted across the leaves)

$$I_G = I(E) - \left[ \frac{1}{2}I(E_1) + \frac{1}{2}I(E_2) \right];$$

2. Information gain weighted by the fraction of documents on each side of the split, as we discussed in class

$$I_G = I(E) - \left[ \frac{N_1}{N}I(E_1) + \frac{N_2}{N}I(E_2) \right],$$

where  $I(E)$  is the information content in the  $N$  examples before the split, and  $I(E_i)$  is the information content of the  $N_i$  examples in the  $i^{th}$  leaf after the split. Method (1) does not deal well with splits that put a lot of examples on one side, and very few on the other. For example, suppose you have evenly distributed data across the two target classes (so  $I(E) = 1$ ):  $N$  examples, and 2 features. Suppose the first feature splits one example off from the other  $N - 1$ , so Method

(2) gives:

$$\begin{aligned} I(E_1) &= 0, \\ I(E_2) &\approx 1, \\ I_G &\approx 1 - \left[ \frac{1}{N} \times 0 + \frac{N-1}{N} \times 1 \right] = \frac{1}{N}. \end{aligned}$$

That is, using Method (2),  $I_G$  is a small value. On the other hand, Method (1) will make  $I_G$  look better than it should be, since it will compute:

$$I_G \approx 1 - \left[ \frac{1}{2} \times 0 + \frac{1}{2} \times 1 \right] = \frac{1}{2}.$$

Use  $I(\{\}) = 1$  (the entropy of the empty set is maximal).

As discussed in class, build each decision tree by maintaining a priority queue of leaves, sorted by the maximum value of the feature (the information gain of the best performing next feature to split on). At each iteration, split the leaf where the split will give the largest information gain (according to each of the measures above), and do the split on the word feature that gives that highest gain. Grow the tree one node at a time, up to 100 internal nodes, by training with the training set only. The point estimate should be the majority class at the leaf (not a probability).

Plot the training and testing accuracy (i.e., percentage of correctly classified articles) after adding each new node. Generate two graphs (one for each method of feature selection) and plot two curves on each (one for training accuracy and one for test accuracy). These graphs should have accuracy as a percentage on the y-axis and number of nodes (1-100) on the x-axis. You may generate them however you wish (e.g. Matplotlib for Python).

### What to hand in:

1. **[20 pts]** Your code, written in python, which must be your own, and you must use the iterative priority queue model. If your code is recursive, or doesn't expand nodes by order of information gain throughout the tree, you will not get these 20pts.
2. **[20 pts]** A printout (or hand drawing) showing two decision trees (one tree for each feature selection method) after adding the first 10 word features. Show the (method-specific) information gain and word feature selected at each internal node. Show the prediction at each leaf. Each printed or drawn tree should have 10 internal nodes and 11 leaves.
3. **[20 pts]** Two graphs (one for each feature selection method) showing the training and testing accuracy as the number of nodes increases from 1 to 100.

## 2 Question 2 [40pts]: Bayesian Networks and Variable Elimination

Suppose you are working for a financial institution and you are asked to implement a fraud detection system. You plan to use the following information:

- When the card holder is travelling abroad, fraudulent transactions are more likely since tourists are prime targets for thieves. More precisely, 1% of transactions are fraudulent when the card holder is travelling, where as only 0.4% of the transactions are fraudulent when they are not travelling. On average, 5% of all transactions happen while the card holder is travelling. If a transaction is fraudulent, then the likelihood of a foreign purchase increases, unless the card holder happens to be travelling. More precisely, when the card holder is not travelling, 10% of the fraudulent transactions are foreign purchases where as only 1% of the legitimate transactions are foreign purchases. On the other hand, when the card holder is travelling, then 90% of the transactions are foreign purchases regardless of the legitimacy of the transactions.
- Purchases made over the internet are more likely to be fraudulent. This is especially true for card holders who don't own any computer. Currently, 60% of the population owns a computer and for those card holders, 1% of their legitimate transactions are done over the internet, however this percentage increases to 2% for fraudulent transactions. For those who don't own any computer, a mere 0.1% of their legitimate transactions is done over the internet, but that number increases to 1.1% for fraudulent transactions. Unfortunately, the credit card company doesn't know whether a card holder owns a computer, however it can usually guess by verifying whether any of the recent transactions involve the purchase of computer related accessories. In any given week, 10% of those who own a computer purchase with their credit card at least one computer related item as opposed to just 0.1% of those who don't own any computer.

- a [10 pts] Construct a Bayes Network that your fraud detection system can use to identify fraudulent transactions.

**What to hand in:** Show the graph defining the network and the Conditional Probability Tables associated with each node in the graph. This network should encode the information stated above. Your network should contain exactly six nodes, corresponding to the following binary random variables:

- *OC* – card holder owns a computer.
- *Fraud* – current transaction is fraudulent.
- *Trav* – card holder is currently travelling.
- *FP* – current transaction is a foreign purchase.
- *IP* – current purchase is an internet purchase.

- *CRP* – a computer related purchase was made in the past week.

The arcs defining your Bayes Network should accurately capture the probabilistic dependencies between these variables.

- b [15 pts] What is the prior probability (i.e., before we search for previous computer related purchases and before we verify whether it is a foreign and/or an internet purchase) that the current transaction is a fraud?

What is the probability that the current transaction is a fraud once we have verified the following 3 facts: (1) it is a foreign transaction; (2) it is not an internet purchase; and (3) the card holder purchased computer related accessories in the past week? (Note: we are describing one query, not three separate queries.)

**What to hand in:** Indicate the two queries (i.e.,  $Pr(variables|evidence)$ ) you used to compute those probabilities. Show the factors computed at each step of variable elimination. For each step, show a formula giving the new factor to be created, e.g.

$$f_7(B) = \sum_A [f_2(B, A) \cdot f_3(A) \cdot f_6(A, B)]$$

$B$	$f_7(B)$
t	0.007
f	0.640

And show the resulting factor (you don't need to show all steps of the calculation). Use the following variable ordering when summing out variables in variable elimination: *Trav, FP, Fraud, IP, OC, CRP*. No marks are awarded for answering the questions using external software.

- c [5 pts] Suppose you are not a very honest employee and you just stole a credit card. You know that the fraud detection system uses the Bayes net designed earlier but you still want to make an important purchase over the internet. What can you do prior to your internet purchase to reduce the risk that the transaction will be rejected as a possible fraud? This question is independent of question 2b.

**What to hand in:** State the action taken and indicate by how much the probability of a fraud gets reduced. Follow the same instructions as for Question 2b to show that the probability decreases. You don't have to replicate all the calculations, only those that change.