PROBLEM SET I

MAT 6480 / STT 6705 - Fall Semester 2019

Your solutions must be in a single zip file titled `ps1.zip`. The zip file should include a single PDF titled `ps1.pdf` and MATLAB or Python scripts as specified. Your homework should be submitted via StudiUM before Thursday, Oct. 3, 2019, at 23:59.

Problem 1

This problem uses the text file `tweets.txt` available on the public course website and on StadiUM. The purpose of this exercise is to write a script called `problem1.m` or `problem1.py` that builds a document-term representations of tweets and then analyzes correlations in them. For this exercise, each of the 188 lines (i.e., tweet) of the text file should be considered as a separate document, while the words in the text file of length at least 5 will act the terms. The text file only contains lower case letter, numbers, and whitespace characters. A word is any sequence of alpha-numeric characters separated by one or more whitespace characters. The script should do the following steps:

1. Read and parse the file so into a cell array (or list) of tweets, where each tweet is itself a cell array (or list) of words. Call this array (or list) `tweets`.
2. Find the ten most frequent terms in all the tweets and store them in a cell array (or list) called `terms`.
3. Verify that the most frequent term is `iphone` and ignore it in the next steps, so the script only considers the next 9 terms
4. Build a 188×9 document-term matrix between tweets and terms and call it `A`.
5. Build a 9×9 correlation matrix between terms, based on `A`, and call it `C`.
6. Build a 9×2 cell array (or list of lists) where the first column contains the terms you used in steps 4-5 in alphabetical order, and the second column contains for each of these terms, its most correlated term (excluding itself) based on `C`. Call this cell array (or list of lists) `pairs`.
7. Print the term pairs in `pairs`.

Include the 9 pairs of printed words in your PDF titled `ps1.pdf`.

Problem 2

Let S^k denote the k-dimension sphere in $\mathbb{R}^{k+1}$ defined

$$S^k = \{x = (x_1, x_2, \ldots, x_k, x_{k+1}) \in \mathbb{R}^{k+1} : \|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_k^2 + x_{k+1}^2} = 1\}.$$

Given a set of points $X \in \mathbb{R}^d$, the set of pairwise distance D is defined

$$D = \{\|x - y\|_2 \in \mathbb{R} : x, y \in X.\}.$$

Write a script called `problem2.m` or `problem2.py` does the following for $k = 1, 2, 3$.

1. Generate 1000 points X_k uniformly at random on $S^k \subset \mathbb{R}^{k+1}$.
2. Compute the set of pairwise distances D_k for X_k.
3. Create a histogram with 25 bins of equal width of D_k

4. Save the histogram as `eqwidth_`k`.png`.
5. Create a histogram with 25 bins each containing an equal number of points of D_k.
6. Save the histogram as `eqpoints_`k`.png`.

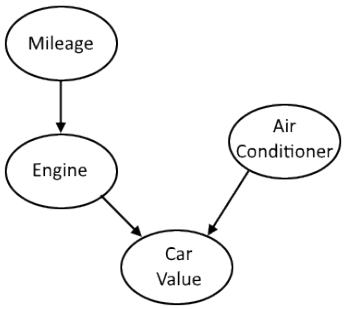
Include the six plots in your PDF titled `ps1.pdf`.

Problem 3

1. Consider the following aggregated dataset with binary attributes:

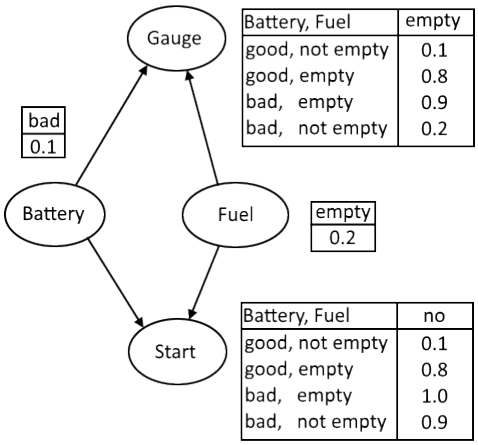
Mileage	Engine	Air Conditioner	frequency(Car Value = Hi)	frequency(Car Value = Low)
Hi	Good	Working	3	4
Hi	Good	Broken	1	2
Hi	Bad	Working	1	5
Hi	Bad	Broken	0	4
Lo	Good	Working	9	0
Lo	Good	Broken	5	1
Lo	Bad	Working	1	2
Lo	Bad	Broken	0	2

   - Complete the CDTs for the following Bayesian belief network:



   - Using this network, compute $\Pr[E = \text{Bad}, AC = \text{Broken}]$, and explain your computation.

2. Consider the following Bayesian belief network and provide detailed computations of the following probabilities:



   - $\Pr[B = \text{good}, F = \text{empty}, G = \text{empty}, S = \text{yes}]$
   - $\Pr[B = \text{bad}, F = \text{empty}, G = \text{not empty}, S = \text{no}]$
   - The probability that the car starts given that the battery is bad

Problem 4

1. Give an example of binary classification with two binary attributes where Gini index prefers one attribute for a (binary) split while information gain prefers the other attribute (*hint:* you can use duplicate entries of each combination of values). Explain how such an example can occur, even though we saw in class that both entropy and Gini index increase and decrease monotonically in the same regions (i.e., $[0, 0.5]$ and $[0.5, 1]$ correspondingly) for binary classification.

2. Consider the setting of $k \geq 2$ classes, nominal attributes, and multiway splits. Prove that information gain of a split is always positive (i.e., entropy never increases by splitting a node). *Hint #1:* You can use Jensen's inequality for the log function to get $\sum_k a_k \log(b_k) / \sum_k a_k \leq \log(\sum_k a_k b_k / \sum_k a_k)$. *Hint #2:* think in terms of joint and conditional probabilities.

Problem 5

1. Given N data points and a real-valued numerical attribute, suggest an algorithm with complexity $O(N \log N)$ for finding the best binary split of this attribute w.r.t. the Gini index. Justify the complexity of the proposed algorithm.

2. Sketch a full decision tree for the parity function over four binary attributes: A, B, C, and D (e.g., the class is $+$ when the number of attributes that are 1 is even, and it is $-$ when this number is odd). The tree should consider a uniform distribution of all possible combinations for these attributes (i.e., $0000, 0001, 0010, 0011, \ldots$). Can this tree be pruned without degrading its classification performances?

3. Consider the following dataset:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Build a decision tree for deciding whether to play tennis or not based on the weather today.

- Your tree construction should be based on Gini index and multiway splits.
- Provide a sketch of the constructed tree in the submitted PDF.
- Provide justifications for the choice of each attribute in the construction process.