CSE 5243: Homework #2

Deadline: **11:59PM on 10/02/2022**.
No late submissions will be accepted.

Instructions.

We currently use a 100-point scale for this homework, but it will take 10% of your final grade.

What you should turn in (**Instructions Different from HW1!**):
(1) For Problem 1-4, please prepare your answers in a single **PDF** file, named as something like HW2-p1-p4.pdf.

(2) For Problem 5, please put your code and all documentations/reports in a folder named as something like HW2-p5.

(3) Put all your files (Problem 1 to 5) in a Zip file named as something like HW2.zip, and submit it to Carmen. See Problem 5 for more info.

Questions?
Please create a post on Carmen discussion areas to get timely help from other students and the instructor. Everyone can benefit from your questions! Please try to avoid directly sending emails to the instructor, unless they are about something personal and irrelevant to the rest of the class.

Problem 1 (6 points).

For the following group of data:
200, 400, 600, 800, 1300, 1500, 2000

(a) Calculate its mean and variance.

(b) Normalize the above group of data by min-max normalization with min = 0 and max = 20.

(c) In z-score normalization, what value should the first number 200 be transformed to?

Problem 2 (14 points).

Given the following table,

X_1	X_2
-3	a
3	b
-4.4	a
6.0	a
-4.0	a
-12.0	b
1.2	a
16.0	b
-16.0	b
13.2	a

assuming that X_1 is discretized into three bins as follows:
$c_1 = (-20, -5]$; $c_2 = (-5, 5]$; $c_3 = (5, 20]$

Answer the following questions:

(a) Construct the contingency table between the discretized X_1 and X_2 attributes. Include the marginal counts.

(b) Compute the χ^2 statistic between them.

Problem 3 (10 points).

Consider the following data matrix D, where rows are data instances and columns are features/attributes:

X_1	X_2
8	-20
0	-1
6	-10
5	-15
2	0
-2	5

(a) Compute the mean of X_1 and X_2 respectively.
(b) Compute the covariance matrix $\sum$ for D.

Problem 4 (50 points).

Assume we get some data from a car insurance company in Table 1, where there are 6 data instances representing 6 people, with 2 attributes (Age and Car) and

1 class label (Risk). Here Age is a continuous attribute. Now we will build decision trees for this data set.

(a) Let us consider a multi-way split for the Car attribute (using its unique values for partition). What is the information gain if we choose the Car attribute to split the root node? (5 points)

(b) Let us consider the binary splits for the Car attribute. Using information gain as the measure, which binary split of the Car attribute is the best at the root node? (5 points)

(c) Between (**a**) and (**b**), which one do you prefer for splitting the root node using the Car attribute? Hint: Consider the GainRatio measure. (5 points)

(d) Now, construct an entire decision tree for the given data set, using information gain as the split point evaluation measure. You can use your calculations or conclusions in (**a-c**). (30 points)

(e) Classify the point (Age=27, Car=Vintage) based on the constructed decision tree in **(d)**. (5 points)

Data Point	Age	Car	Risk
x_1	25	Sports	L
x_2	20	Vintage	H
x_3	25	Sports	L
x_4	45	SUV	H
x_5	20	Sports	H
x_6	25	SUV	H

Table 1: Data for Problem 4. *Age* is numeric and *Car* is categorical. *Risk* gives the class label for each point: high (H) or low (L).

Problem 5 (20 points). Programming Task on Real Text Data Preprocessing.

This assignment is the first part of a longer-term project. The objective is to give you the experience of preprocessing real data and preparing it for future tasks such as automated classification.

- **Data**: Sentiment Labelled Sentences Data Set, which contains sentences labelled with positive or negative sentiment. It can be downloaded here `http://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences`. Read their readme.txt file for detailed information. There are three subsets respectively from IMDB, Amazon and Yelp. Please merge them as a single dataset, which should contain 3,000 sentences in total.

- **Data Format**: Each data file is .txt where each row has two columns: sentence body and sentence label. For example, one sample sentence is “Very little music or anything to speak of. 0”, where the first column “Very little music or anything to speak of.” is the content of a sentence while the second column “0” is its sentiment label (1 means positive; 0 means negative).

- **Task**: In this assignment, your task is to construct a feature vector for each sentence in the data set. For now, please use the frequency of words in the sentence body to construct a feature vector. For example, if there are totally M sentences and N words in the dataset, you will construct a $M \times N$ matrix D, where $D_{i,j}$ means the count of word j in sentence i. Hint: You first need to segment/tokenize a sentence to get a collection of words in it. After that, it is up to you whether to do stemming (e.g., “likes” and “liked” are stemmed to “like”) or simply keep the original words.

- **Programming Requirement**: ***You are required to use Python as programming language. If you want to use other programming languages, please talk to the TA and make sure she can run your code successfully in the stdlinux environment***. Note – you are encouraged to implement standalone code (Python suggested) for this project (including upcoming programming related tasks), and it is also OK to leverage free software or packages such as NLTK (Natural Language Toolkit) as long as you can concisely and precisely explain how you used them in your report or README file.

What You Should Turn in:

You are expected to turn in the following files:

- All source code. Please make it easy to run your code. You do not have to include the raw dataset.

- A detailed README file that contains all the information about the folder, e.g., what files are in the folder, how to run your program, how to interpret the output of your code (e.g., you can sample 5 input sentences and show their non-zero features after preprocessing, respectively. This will help TA and yourself understand you did the right preprocessing), etc.

- A short report (no more than 1 page in 12 pt. font), describing the approach/procedure you took to construct a feature vector, and where applicable the rationale for doing it (name this file **report1.pdf** or some such thing).

Detailing what you did is very important even if it did not work. Describe any difficulties you may have encountered and any assumptions you are making. Important: **You need to clearly state what you filtered and what you did not filter from your preprocessing**.

How to submit all your files to Carmen:

After you finish all the problems, you just need to upload a single .zip file to Carmen, as we mentioned in Instructions.

You don't have to do the following, but just FYI: To open carmen from stdlinux,

- Login to stdlinux using CSE remote access. Check here for *remote access*, if needed: `https://cse.osu.edu/computing-services/resources/remote-access`.
- Open terminal
- Type "firefox". Click enter
- Navigate to carmen.osu.edu

If it is necessary to transfer files from Windows/Mac environment to stdlinux, one can make use of SCP and SFTP protocols to achieve so. This site[1] can provide additional details. Contact CSE Help Desk if further help is needed setting up File transfer clients.

Necessary Details:
a. Host Name: stdlinux.cse.ohio-state.edu
b. Port Number: 22
c. User Name: osu user-id
d. Password: OSU password

Acknowledgement. The dataset was originally used in the following paper:
[1] "From Group to Individual Labels using Deep Features", Kotzias et al., SIGKDD 2015.

[1] https://u.osu.edu/floss/documentation/using-scp-sftp-for-ohio-state-cse-students-to-sync-projects-and-remotely-access-your-files/