CSCN8000 – Artificial Intelligence Algorithms and Mathematics – Spring 2024

Lab 2

This lab provides insight into the concepts of probability, statistics and linear equations.

Theoretical Part A [10 Points]

In this part, you will be utilizing the Naïve Bayes approach in order to generate predictions from the housing dataset shown below. The following dataset represents the Price of a house given it's location and size as input features.

Location (L)	Size (S)	Price (Label)
Urban	Large	Expensive
Suburban	Medium	Affordable
Rural	Small	Cheap
Urban	Medium	Affordable
Suburban	Large	Expensive
Rural	Medium	Affordable
Urban	Small	Cheap
Suburban	Small	Cheap
Rural	Large	Expensive
Urban	Large	Expensive

Using the Naive Bayes approach done in class, calculate the following **and** determine the predicted price (Expensive/Affordable/Cheap). Make sure to include ***All the steps and calculations*** needed to reach the final answer. Only 2 Points will be granted for those who submit the final answer without steps.

- $P(Price|L = Urban,\ S = Medium)$

Theoretical Part B [10 Points]

Here is a system of linear equations (or linear system) with three equations and three unknown variables:

$$\begin{aligned} 4x_1 - 3x_2 + x_3 &= -10, \\ 2x_1 + x_2 + 3x_3 &= 0, \\ -x_1 + 2x_2 - 5x_3 &= 17, \end{aligned}$$

Solve this system of linear equations using Gaussian Elimination to find the values of the variables x_1 , x_2, x_3 such that all of its equations are simultaneously satisfied. Make sure to include ***All*** the steps and calculations needed to reach the final answer. No points will be granted for those who submit the final answer without steps.

Note: to validate if your final answer is correct or not, you can use Python along with the np.linalg.solve() function to get the answer and cross-check it with your derived answer. Again, this is only for you to validate the correctness of the steps, however, submitting only the final answer even if correct will not be granted any points.

Deliverables for Theoretical Parts A,B

1. Those parts are written not coding sections. You're free to provide you answers to those parts as handwritten scanned papers or electronically written documents (i.e. Word Document/Latex). Please make sure your submissions are neat, organized, and easy to read or it will be hard to fairly grade your assignment.
2. Submit a final scanned pdf named as "[Full Name]_[Student ID]_[Section Number]_Theoretical_Lab2.pdf"

Practical Part A [15 Points]

Use the **Lab2_dataset.csv** provided. Your target in this part will be classify texts into Spam vs Not Spam using the Naive Bayes algorithms and comparing them to another type of classification model and comment on the results.

Preprocessing [5 Points]

- Load the dataset
- Use the CountVectorizer function in sklearn to transform the "text" feature to a vector representation of a predetermined size.
- Split the dataset into training and testing

Model Training and Evaluation [10 Points]

- Train the Sklearn RandomForestClassifier model on the training dataset and evaluate on the test set
- Train and evaluate also on the Gaussian and Multinomial Naiive Bayes Classifiers
- Compare between the performance of all models and comment on the reasons behind the differences seen between the three models.
- ***Note that the* RandomForestClassifier *model doesn't make the same assumptions as the other Naiive Bayes models***

Practical Part B [10 Points]

Use the **AB_NYC_2019.csv** dataset for this part.

Tasks

- Remove outliers based on price per night for a given apartment/home.
- Compare the Z-score approach and the whiskers approach in terms of who is better to remove the outliers in this case.

The task is to come up with a clean dataset that does not have outliers showcasing all the possibilities

Organization Criteria for All Parts (5 Points)

1. (5 points) Provide an organized full submission of all parts. For the practical parts, make sure your notebook has clear sections, markdown comments and printed outputs. Make sure no errors are printed in the final submission.

Deliverables for Practical Parts A,B

1. Include all your findings and task solutions in one Jupyter notebook (.ipynb) that shows all the printed cell outputs. Prepare a html version (.html) of the notebook file. Both files should be named as follows: [Full Name]_[Student ID]_[Section Number]_Practical_Lab2.[html/ipynb].
2. Submit both .html and .ipynb files on eConestoga in the Lab 2 under the Assignments section.

Late Submission Policy

1. 10% of the grade will be deducted for each day that your submission is late.