Course: INT 93S
Quarter: Summer 2017
Assigned: 13 July 2017
Due: 9:00 18 July 2017

Lab 5: Unsupervised Learning

Collaboration Guidelines: Code and written analysis can be discussed with your lab partner, but must be written separately, (unless you are working on the same code on the same computer together at the same time). Your lab partner is the only student who can look at your code. You can discuss the lab report but it must be written separately.

1 point will be taken off for not including a readme, improper lab format, and/or folder named improperly.

Turn-in:

1. Create a readme file titled "lastname_firstname_readme.txt" which includes:
    a. First name and last name of you and your partner
    b. How much of the lab you finished in class
    c. Any references you used
    d. Anything you would like the professor and TA to know.
2. Put your code and report in a folder named "lastname_firstname_lab5"
3. Zip the folder and upload it to Gauchospace

Report Format: *(1pt)*

1. In a document file (word, google docs, etc), write up an analysis for the lab.
2. Include your name, the date and the title of the lab.
3. Include the Analysis from 1 and 2, as well as the final Decision Tree Diagram.
4. Export Lab report as a pdf saved as "lastname_firstname_lab5_report.pdf"
5. Follow turn-in instructions, including the readme and proper folder/file names

Part 1:

Write an application to calculate Naive Bayes to predict if a car is "acceptable" or "unacceptable" from scratch for dataset carCat.csv. See carMod.txt for more info on the dataset. *(4pts)*

1. You may use naiveBayesExampleCode.py as starter code. Name your python file "lastname_firstname_lab5p1.py"
2. Analysis:
   What is the probability of it being an acceptable car in these situations? *(1pt)*
    a. For each attribute, which label predicts the highest acceptability?
    b. Calculate the below probabilities:
        i. P(acceptable | buying:high, maint:v-high, doors:4)
        ii. P(acceptable | buying:low, maint:med, safety:med)
        iii. P(acceptable | buying:high, maint:v-high, safety:low)

iv. P(acceptable | doors:5more, persons:more, lug_boot:big)
v. P(acceptable | doors:2, persons:2, lug_boot:small)

Part 2:

Use scikit learn to create different models for the dataset carVal_train.csv and test the resulting model on carVal_test.csv

1. Name your python file "lastname_firstname_lab5p2.py"
2. Using scikit learn to create a Gaussian Naive Bayes Model based on carVal_train.csv and test it's accuracy on carVal_test.csv. *(1pt)*
3. Using scikit learn to create a Decision Tree based on carVal_train.csv and test it's accuracy on carVal_test.csv. *(1pt)*
    a. Use "entropy" as your criterion.
    b. Use tree.export_graphviz() to create a dot file and use http://www.webgraphviz.com/ to view it.
    c. Test at least 4 different values for "max_depth," and 4 different values "min_samples_split" and see how it affects the accuracy and layout of the tree.
    d. Redraw the best tree solution on your computer or by hand based on your dot diagram *(1pt)*
        1. Label with attributes and words instead of index values
        2. Include the entropy, the number of samples and the distribution per node.
2. Analysis: *(1pt)*

    In paragraph form answer the following questions.
    a. How accurate are your models? Which works better?
    b. Why does one model work better than the other?
    c. Explain your choices for the Decision Tree Model:
        i. What parameters did you use for the algorithm?
        ii. How did the parameters affect the shape of the tree?
        iii. How did the parameters affect the accuracy?
        iv. How did you avoid overfitting?

Extra Credit 1: *(.5 pt)*

1. Calculate the accuracy of your Naive Bayes algorithm and compare it to the results of Part 2.

Extra Credit 2: *(1pt)*

1. Find a categorical dataset related to your research or your partners research project. Evaluate it using scikit learn create a Decision Tree and Gaussian Naive Bayes Model.
2. Write a paragraph analyzing your results:
    a. How accurate are the models?
    b. Would you use one of these models to explain your dataset?