DSCI-549: Introduction to Computational Thinking and Data Science

USC Viterbi School of Engineering

Homework 3

Communication

Please contact TA **Yashika Goyal** (yashikag@usc.edu) for this homework and include "**DSCI549**" in the subject line of your email.

Assignment

The purpose of this homework is to use a variety of machine learning algorithms, such as decision tree classifiers, and understand their limitations and performance. You must complete this assignment using the Jupyter Notebook. Answer the questions in a separate word document, along with relevant screenshots. Submit the homework to D2L.

For this assignment, you will be using three notebooks. You can access them here:

https://github.com/doctorningwangusc/DSCI549/tree/main/Homework%20Assignments/Assignment3_MachineLearning

- Notebook1_Decision_Tree_Classifier.ipynb
- Notebook2_Naive_Bayes_Classifier.ipynb
- Notebook3_Overfitting_and_Underfitting.ipynb

For this assignment, you will be using two datasets described in the following links:

- lenses.csv
- iris.csv

Note: The iris data has continuous values, the contact lenses are discrete.

1. Using a decision tree classifier (35 points)

    a. Run the workflow contained in Notebook1_Decision_Tree_Classifier.ipynb using the contact lenses dataset. Include the resulting visualization here.
    b. You will evaluate the accuracy of the decision tree classifier using k-fold cross-validation. Run the cross-validation using a number of folds k = 3. Using your understanding of k-fold cross-validation, describe in your own words the testing/training visualization. What is the accuracy of the classifier?
    c. Using the trained classifier, make a prediction for a new patient with the following characteristics: *"young, myope, astigmatic, reduced"*. What features did you input and what result did you get?
    d. Rerun the cross-validation for the number of folds k = 7 [Include appropriate screenshot(s)]. Can you explain the warning in the context of your dataset?
    e. Rerun the workflow for the iris dataset and include a screenshot of the resulting decision tree. Apply an appropriate number of folds for this dataset and tell what is the accuracy of the classifier for this dataset?

2. Comparing different Naïve Bayes classifiers (35 points)

   a. Using the workflow contained in Notebook2_Naive_Bayes_Classifier.ipynb you will be training/validating two types of Naïve Bayes Classifiers: the Gaussian and Multinomial Naïve Bayes Classifiers. Which dataset is most appropriate for the Gaussian Naïve Bayes Classifier? Why? Which dataset is most appropriate for the Multinomial Naïve Bayes Classifier? Why?
   b. Run the classifiers with the appropriate data and include screenshots. What are their accuracy scores?

3. Analyzing overfitting (30 points)

   a. Run the workflow contained in Notebook3_Overfitting_and_Underfitting.ipynb, for various degrees for the polynomial fit. Provide screenshot(s) of all the different degrees for the polynomial you have tested. [Try with at least 4 different degrees for the polynomial]
   b. Provide an example of overfitting and underfitting of the data. Provide screenshots of the visualization to support your statement.

IMPORTANT NOTES

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in the Student Handbook https://policy.usc.edu/studenthandbook/. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, https://policy.usc.edu/research-and-scholarship-misconduct/.

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the American Language Institute http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students.

For more information, see the class syllabus and the USC web site.