CS 455/595a – Homework #3: KNN and Regression Models

Classification and Regression Activities

Each student must submit their own work. The first two tasks can be completed collaboratively or individually. The CS 595 only task must be completed individually.

Submit a zip file with the requested deliverables. Your zip file should include your last name. Each source code and/or summary file submitted should also include your name.

Classification Tasks [50 points] – CS 455 & 595 - Individually or Collaborate with up to one partner

Deliverable: Source Code and Summary

For this problem, you will develop multi-class classification models to analyze the Wine dataset provided by SKLearn: https://scikit-learn.org/stable/datasets/index.html#wine-dataset.

For this task, you will create a Jupyter notebook or a .py Python script to perform the following analysis using Python, SKLearn, and other relevant libraries.

You may use the classification examples provided by Dr. Stansbury on Github for ideas of metrics, analysis, and code snippets to inform your implementation. Your notebook must do the following:

- Import the wine data set
- Analyze the wine data set including:
    - Output the shape of the data set
    - Describe the minimum, maximum, and average values for each of the features and the target labels.
    - Implement a scatter matrix to determine if any features strongly correlate.
- Pre-process the data as necessary to ensure that all values can be processed by a classification model.
- Develop, train, and demonstrate performance by plotting the ROC curve, showing the confusion matrix, showing the performance (accuracy, precision, recall, f1 score) for the following configurations:
    - K-Nearest Neighbors
    - Logistic Regression

You are not being asked to implement the algorithms. Use the model libraries in SKLearn to build your classifier models.

When you are finished, write a brief summary of the work you did. If you are using a Jupyter notebook, your writeup can accompany your code. If not, you must write up your summary in a separate document with screenshots as necessary to demonstrate your work.

Regression Tasks [50 points] – CS 455 & 595 - Individually or Collaborate with up to one partner (please clearly identify your partner)

Deliverable: Source Code and Summary

For this problem, you will use the Diabetes data set from sklearn. It is documented at: https://scikit-learn.org/stable/datasets/index.html#diabetes-dataset.

For this task, you will create a Jupyter notebook or a .py Python script to perform the following analysis using Python, SKLearn, and other relevant libraries.

You may use the regression examples provided by Dr. Stansbury on Github for ideas of metrics, analysis, and code snippets to inform your implementation. Your notebook must do the following:

- Import the diabetes data set
- Analyze the diabetes data set including:
    - Output the shape of the data set
    - Describe the minimum, maximum, and average values for each of the features and the target values.
    - Implement a scatter matrix to determine if any features strongly correlate.
- Pre-process the data as necessary to ensure that all values can be processed using a regression model.
- Develop, train, and demonstrate performance by plotting the learning curves for the following configurations:
    - A linear regression model
    - A polynomial regression model
    - A model will regularizations implemented (linear or polynomial)

You are not being asked to implement the algorithms. Use the model libraries in SKLearn to build your classifier models.

When you are finished, write a brief summary of the work you did. If you are using a Jupyter notebook, your writeup can accompany your code. If not, you must write up your summary in a separate document with screenshots as necessary to demonstrate your work.

CS 595a Research Task [50 points] – Complete Individually

Select one algorithms used within this assignment and through some research and/or analysis answer the following questions in a brief 1-page report.

1. What is the model?
2. Briefly describe the model. If it is related to one of the models already covered, be sure to discuss how they interrelate.
3. What types of problems does the model typically solve? Give at least two examples of problems where the algorithm has been applied and cite your sources.

4. What is the computational complexity of the algorithm? Answer in terms of run-time performance using big-O notation considering the applicable parameters for the algorithm's performance including number of features, number of instances, etc.?
5. Does the model require any pre-processing of the data before its use?
6. Does the model support online learning? Out-of-core training?
7. What is the origin of the algorithm? Reference an early paper that makes reference to the model and/or the family of models it serves.

The purpose of this exercise is to give you practice in looking deeper into the models used and how research in model development and applications influence our understanding of the algorithm.

Citations and references must be in IEEE style for full credit.

BE SURE TO INCLUDE YOUR NAME IN ALL OF THE FILES YOU SUBMIT THAT ARE TO BE GRADED.