CSE 587: Exam-1

Due March 08, 2022 (Tuesday) 05.00 PM

Notes: The exam must be done individually, with no discussion or help with others. Breaking this rule will result in an automatic 0 grade.

PART A (30 Points)

1. Explain overfitting and underfitting. What evaluation metrics do we use to check if the regression model fits the data well? **(5 points)**
2. Differentiate Correlation and Regression. What is R-squared? Discuss the drawback of R-squared in the case of multiple linear regression. What is Adjusted R-squared? **(5 points)**
3. Explain precision and recall using the confusion matrix. Discuss with examples when precision and recall may consider as more useful classification evaluation metrics over accuracy. **(5 points)**
4. What is the advantage of K-Means Clustering over Hierarchical Clustering? What are Manhattan Distance and Euclidean Distance in Clustering? Why may we consider different distance measures for clustering? **(5 points)**
5. What is feature scaling? Show an example of why feature scaling is crucial? What are the two standard feature scaling techniques? **(5 points)**
6. Discuss collaborative and content-based recommendation techniques with examples. **(5 points)**

PART B (25 Points)

A data set named "DATA.csv" is uploaded to the Exam-1 folder. This dataset represents traffic violation information from all electronic traffic violations issued in the US Counties. This data is de-identified; any information that can be used to uniquely identify the vehicle, the vehicle owner, or the officer issuing the violation is removed. The target variable in this data set is the *"Violation Type."*

7. Develop classification models using *K-Nearest Neighbour, Random Forest, and Support Vector Machine*. Use domain knowledge to select your features to train the model with justification. Take advantage of data preprocessing techniques, where applicable, to prepare the data to fit into the machine learning algorithms. Compare these three models in the lens of accuracy, sensitivity, and specificity when deployed on an unseen test set. Also, indicate how generalizable each model is. Which model will you suggest to deploy? Discuss your rationales behind this selection.

PART C (20 Points)

Survey the paper on **"Exascale Computing and Big Data."** This paper is authored by Daniel A. Reed and Jack Dongarra. (Use this link to download the paper: Link)

8. Write a summary of this paper. Answer the following questions in your summary. **[Suggested word count: 250-300 words]**
    - Discuss the context and overarching motivation of this paper.
    - Discuss the scientific and engineering opportunities with the rise of big data analytics.
    - Discuss the technical challenges towards high-end computing examined in this article.