Shahid Beheshti University

Machine Learning Fundamentals

3red Assignment – June 3, 2024

Due date: **June 6, 2024**

Hello everyone, I trust you are all in good health and spirits. This is the third assignment for our Machine Learning course. The deadline for this assignment is **June 12, 2024**. I encourage all students to adhere to this timeline for submissions. Should you have any questions concerning the exercises, please feel free to reach out.

Part 1

a) Explain and compare four different kernels used in Support Vector Machines (SVM) (namely, linear, polynomial, RBF and Sigmoid)

b) Compare and contrast CatBoost and LightGBM with XGBoost. Discuss the unique features and optimizations of each algorithm, and their impact on performance, especially in handling categorical features and large datasets.

c) Compare stratified k-fold cross-validation and regular k-fold cross-validation. When do we use stratified k-fold cross-validation?

d) Discuss the challenges of validating clustering results in the absence of ground truth labels. Explore clustering metrics and their acceptable vales and using domain-specific knowledge.

e) Explain how you would determine the optimal number of clusters (K) for a given dataset. Compare different methods, such as the Elbow method, Gap statistic, and Silhouette analysis.

f) What's the impact of the choice of different k in k-fold cross-validation on bias-variance tradeoff in model evaluation?

g) Describe the out-of-bag (OOB) error estimation in bagging. How is it computed, and why is it useful?

h) Explore the application of PCA in the field of genomics. Discuss how PCA can be used to identify patterns in high-dimensional genetic data.

i) Explore the application of SVM algorithms in anomaly detection.

Part 2

In this section, you will perform an emotion detection task using a single-label collection of Persian texts categorized into five emotional classes: happiness, sadness, anger, fear, and others.

- You are expected to perform **data cleaning** and **feature engineering** on the dataset. Subsequently, you should construct a brief report detailing your methodology and ideas within your Jupyter notebook or in a separate document.

- Explore tree-based classification algorithms. test various models, adjusting hyperparameters to optimize performance. Utilize model pruning to streamline your model if necessary. Provide a summary of your findings and interpret your results in your report.

- For the final evaluation, you are encouraged to use whichever classical model you find most effective in achieving optimal performance.

- You must evaluate your model using **appropriate metrics** on a validation set and provide an interpretation of your metric scores within your report. You can use cross k-fold cross-validation or stratified k-fold cross-validation to provide a more reliable estimate of your model performance.

- A test set without true labels will be provided for you to make predictions using your model during the main evaluation. Additionally, please ensure that the inference process is well-documented in your notebook, allowing us to reproduce and verify your results.

- Constructing a comprehensive pipeline that encompasses data preprocessing, model training, and inference will earn you **bonus points**. This pipeline should be clearly outlined in your notebook.

- Please notice that achieving an accuracy level above a **specified threshold** on the test set will result in full marks for that segment. It is essential to understand that grading is **not competitive**.

Plagiarism will not be tolerated. Homework submissions will be cross-checked against other students' submissions. Additionally, **the use of AI to fully generate answers or code** for assignments is strictly forbidden.