Homework #4

You must sign your name at the end of your submission

1. Consider the below dataset. Applying Knowledge to the field of Medical Science and making the task of Physician easy is the main purpose of this dataset. This dataset has 132 parameters on which 42 different types of diseases can be predicted. Based on the dataset, now answer the following questions:(code + theory)
   1. What type of problem is the given dataset? Explain with examples. 03
   2. What are some popular algorithms we have introduced in our class that can be used for training a classification model on the given dataset? Compare their pros and cons. [Show at least 3 models]

06

* 1. How do you evaluate the performance of a classification model? Discuss different evaluation metrics and explain their significance. 07

1. Consider the dataset of Question 1 and answer the following questions based on the dataset:(code+theory)
   1. What is hyperparameter tuning? How can you use it to improve the performance of a classification model on the given dataset? 07
   2. Can you implement a neural network model to classify the given dataset? Discuss the architecture and hyperparameters of the model. 06
   3. How can you handle class imbalance in the given dataset? Discuss different techniques and their effectiveness. 04
2. You are given a dataset containing information about customer churn for a telecommunications company. The dataset has a mixture of continuous and categorical variables. You have been asked to build a logistic regression model to predict which customers are likely to churn. (18 points)
   1. Explain what is regularization in machine learning? Why is it important? 03
   2. Discuss the L1 and L2 regularization techniques. What is the difference between them? 05

iii) Implement a logistic regression model on a given dataset and tune the regularization hyperparameter. Discuss the effect of regularization on the model performance. 07

1. Explain the bias-variance tradeoff in machine learning. How can regularization help to balance it? 04
2. Suppose you have a dataset consisting of the following 8 points in a two-dimensional space: {(2, 10), (2, 5), (8, 4), (5, 8), (7, 5), (6, 4), (1, 2), (4, 9)}. (20 points) [4X5=20]
   1. Use the K-Means Algorithm with K=2 to cluster the data points into two clusters. Show your work step-by-step, including the initial random selection of cluster centers, the calculation of the distances between data points and cluster centers, and the reassignment of data points to clusters.
   2. Calculate the Cluster Purity of the resulting clusters from part a), using the ground truth labels {(2, 10), (2, 5), (1, 2), (4, 9)} as belonging to Cluster 1 and the rest belonging to

Cluster 2.

* 1. Use the Elbow Method to determine the optimal number of clusters for the dataset from part a). Plot the Within-Cluster Sum of Squares (WCSS) for K=1 to K=7 and identify the "elbow" point where the rate of decrease in WCSS slows down significantly.
  2. Use Density-Based Clustering with a radius of 2 and a minimum density of 3 to cluster the dataset from part a). Show the resulting clusters and explain the advantages and disadvantages of Density-Based Clustering compared to K-Means.

Note: For part a), you can use any method to select the initial cluster centers (e.g., randomly, based on domain knowledge, etc.). However, you must show all the steps involved in the

K-Means Algorithm, including the iterative reassignment of data points to clusters and the recalculation of cluster centers until convergence.

1. Suppose you are working as a data scientist for a financial services company. You are tasked with building a machine learning model to predict whether a customer will default on their loan payments or not. You have been provided with a dataset containing information about past customers, including their credit history, income, debt, and other demographic information. (20 points)
   1. Outline the steps you would take to review the data before building the model. What are some common issues you might encounter, and how would you address them? Please be specific.
   2. After reviewing the data, you decide to use a logistic regression model to predict loan defaults. You train the model on the dataset and evaluate its performance using

cross-validation. However, you find that the model is performing poorly and is not able to predict defaults accurately. What steps would you take to debug the model and improve its performance?

* 1. In addition to the logistic regression model, you also decide to try a decision tree model to predict defaults. After training the model on the dataset, you review its performance and find that it is much better than the logistic regression model. However, upon closer inspection, you realize that the decision tree is overfitting the data and is not able to generalize well to new data. What steps would you take to address the overfitting and improve the generalization performance of the model?
  2. Suppose you are given a new dataset that contains additional features that were not present in the original dataset. What steps would you take to review the new data and incorporate the new features into your model? What are some potential issues that you might encounter, and how would you address them?
  3. In addition to the new features, you also discover that the dataset contains some missing values and outliers. What steps would you take to handle missing values and outliers in the data, and how might these steps affect the performance of your model?

1. Write a short essay discussing the importance of reviewing data and debugging models in the machine learning workflow. Provide examples of how failure to review data or debug models can lead to poor performance and inaccurate predictions, and discuss how these issues can be prevented. 8

I acknowledge and accept the Honor Code.

(Signed) Vishwam Shah

If you are not printing this document out, please type your initials above.