COMP 5600/6600

Assignment 2 – Basic Machine Learning
Due: 10/14/2024 11:59 pm
Submission: On Canvas

Overview:

In this assignment, you will be implementing and analyzing three different machine learning algorithms. Each algorithm will focus on a different problem scenario that can be solved by different learning methods. You must implement all algorithms from scratch using Python and you can use any external libraries such as Sci-kit learn. You can use other libraries for data loading and pre-processing as necessary.

[Problem 1] [25 Points]

You are tasked with developing models to predict customer churn for a subscription-based service. Using the provided dataset, your goal is to build two classification models: one using Logistic Regression and the other using Naive Bayes. You will compare their performance, interpret the results, and provide insights into customer churn based on your findings. You will use provided the Telco Customer Churn dataset, which contains customer information such as demographic details, account features, and whether the customer has churned. Your target variable is "Churn," indicating whether a customer has left the service.

Ensure you follow the below instructions:

- Evaluate both models using the following metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.
- Perform 5-fold cross-validation on both models and report the averaged results.
- If there are any missing values (there will be!), fill them in during a pre-processing step using two of the three common strategies outlined below. Do this for the entire dataset!
    - Use the most common value in the dataset that has a value for this feature/attribute
    - Use a default value to fill in for missing values. It can be anything.
    - Drop that feature all together and use only features that have values for all data points.
- Scale or normalize numerical features if required.
- Ensure that your IPython notebook has text files that has the following details:

- Discuss your outcomes from using your chosen preprocessing steps to handle missing data
- Compare the performance of both models and discuss their strengths and weaknesses. Which model is more suited for this dataset and why?
- Insights gained from your experiments.

[Problem 2] **[25 Points]**

In this question, you will be using k-means to perform image compression. Implement a naïve version of the k-means algorithm based on your understanding. Your code must take the number of clusters k as input and perform k-means clustering on the given image (test_image.png). Once the algorithm finishes running, the cluster centroids represent the top-k common colors in the image. Iterate through each pixel in the image and assign the closest color to each pixel. Save and visualize the resulting image. For reading and writing images, you can use OpenCV, which is an open-source computer vision toolkit. The following code will load the image into a NumPy array. You can use this as input to your K-Means algorithm.

```
import cv2
img = cv2.imread('test_image.png')
height, width, channels = np.shape(img)
for i in width:
for j in height:
pixel = img[j][i] # Read the pixel at location (i,j)
img[j][i] = newValue # Assign a new value to the pixel
```

Experiment with different values of k and briefly describe your thoughts about which value works best for this problem. You can use plots, error bars, etc. to support your conclusions.

Deliverables:

A single IPython Notebook that contains your code and report. You can use the text cells to write your report and embed any plots, illustrations, and/or images that you need to support your claims.