**CITS5508 Machine Learning**

Mid-Semester Test

Student Name: Varun Jain

Student Number: 21963986

Due: Friday 26th March 2021

**Question 1**

For a multiclass classification problem, we are going to use the Support Vector Machine (SVM) classifier to train the dataset. There are two distinct methods we can use: one-verse-all and one verse one.

Step 1: Data Preparation

1. Using the panda’s library, I would load the data into the python script. I would return a Pandas Data frame object containing all the data.
2. Using pandas functions such as head(), info(), describe() to gain insight about the structure, the description such as the total number of rows, the datatype associated with each column and a summary of the numerical attributes return the std, mean and quartile range.
3. Split the data
4. Separate the dependent and independent class.

Step 2: Data Cleaning

1. Remove any irrelevant columns that are not necessary for the computation.
2. For data inconsistency in numerical attributes columns, we can compute the following three options: remove the corresponding row of the inconsistent data, remove the entire numerical attribute from the dataset or replace the missing value to some value such as the mean, median, zero and so on. I would
3. For data inconsistency in the categorial attribute columns, I would first observe the categories in each class, and ensure that each class have proper naming conventions in the sense that

Step 3: Data Transformation

1. For the independent categorial columns, I would use the OneHotEncoder() function from Sklearn’s pre-processing library to transform the categorial attribute to a binary attribute per category in a class. This method is called one hot encoding.
2. Apply feature scaling to the data to standardise as Machine Learning Algorithms do not perform well with unscaled numerical data. I would use the Standardise() method to rescale the data to have a mean of 0 and standard deviation on 1.

Step 4: Train the Model

1. Perform multiclass classification using multiple binary classfiers. To train the binary classifier, we will use one-verse-one approach with a Support Vector Machine (SVM) classifier.

**Question 2**

In order to compute the average precision, we have to compute the precision values for the 3 classes individually: dog, cat and fish. We know that Precision = . Figure 1, shows two distinct colours: red and orange. The orange marker represents the trues positives and the red marker marks all the classes classified correctly and incorrectly, so the true positives and false positives, retrospectively. In reference to diagram (a), we can see that out of the 23 dogs that got classified, only 16 of them got classified correctly, and the remaining 7 classes got classified as false positives. So, to compute the precision for the 3-class classification problem, we do the following:

Precision = .

We compute the same for the remaining two classes: cat and fish.

Diagram

Description automatically generated with low confidence

Diagram

Description automatically generated

To compute the average precision of the three classes, we take the summation of the precision values and divide it by the total number of classes, which in this instance is three.

Diagram

Description automatically generated with medium confidence

To compute the average recall, we follow similar steps that was shown for average precision. Now instead, we calculate the recall value of each of the classes individually. Since we know that, recall = , we take the total number of classes that was correctly predicted divided by the sum of the total number of classes that was correctly and incorrectly predicted. We take the sum of the recall values and divide by the total number of classes in the confusion matrix, which in this case is three.

**Question 3**

1. When the alpha coefficient it too large, the slope of the line will flatten the further the alpha value increases, meaning there is a constraint on the magnitude of the coefficient. Therefore, high alpha value is subject to underfitting.

(ii)

(iii)

**Question 4**

**Question 5**