Minor Project 2 – Cognitive Application

(Ayush Goyal, Manipal Institute of Technology, CSE,aiqqia.ag@gmail.com)

Theory Questions

1. **When should you use classification over regression?**

Classification is about identifying group membership while regression technique involves predicting a response. Both techniques are related to prediction, where classification predicts the belonging to a class whereas regression predicts the value from a continuous set. Classification technique is preferred over regression when the results of the model need to return the belongingness of data points in a dataset to specific explicit categories. Classification is used when the output variable is a category such as "red" or "blue", "spam" or "not spam". It is used to draw a conclusion from observed values. Differently from, regression which is used when the output variable is a real or continuous value like "age", "salary", etc. When we must identify the class, the data belongs to we use classification over regression. Like when you must identify whether a name is male or female instead of finding out how they are correlated with the person.

2. **How do you deal with the class imbalance in a classification problem?**

Data are said to suffer the Class Imbalance Problem when the class distributions are highly imbalanced. In this context, many classification learning algorithms have low predictive accuracy for the infrequent class. Cost-sensitive learning is a common approach to solve this problem. Some other effective solutions are as follows:

a. Change the performance metric
b. Change the algorithm
c. Resampling Techniques – Oversample minority class
d. Resampling Techniques – Under sample majority class
e. Generate synthetic samples

These are just some of the many possible methods to try when dealing with imbalanced datasets, and not an exhaustive list. Some other methods to consider are collecting more data or choosing different resampling ratios — we don't have to have exactly a 1:1 ratio.

We should always try several approaches and then decide which is best for our problem.

3. **What is a confusion matrix and why do you need it?**

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if we have an unequal number of observations in each class or if we have more than two classes in our dataset. Calculating a confusion matrix can give us a better idea of what our classification model is getting right and what types of errors it is making.

The main problem with classification accuracy is that it hides the detail we need to better understand the performance of our classification model. Classification accuracy can hide the detail we need to diagnose the performance of our model. But thankfully we can tease apart this detail by using a confusion matrix.

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. *The confusion matrix shows the ways in which our classification model is confused when it makes predictions.* It gives us an insight not only into the errors being made by our classifier but more importantly the types of errors that are being made. It is this breakdown that overcomes the limitation of using classification accuracy alone.

4. What is the difference between sigmoid and soft max function?

Sigmoid Function acts as an activation function in machine learning which is used to add non-linearity in a machine learning model, in simple words it decides which value to pass as output and what not to pass, there are mainly 7 types of Activation Functions which are used in machine learning and deep learning.

The Soft max function is used as the activation function in the output layer of neural network models that predict a multinomial probability distribution. That is, soft max is used as the activation function for multi-class classification problems where class membership is required on more than two class labels.

The sigmoid function is used for the two-class logistic regression, whereas the soft max function is used for the multiclass logistic regression.

If we have a multi-label classification problem that is there is more than one "right answer" in which the outputs are not mutually exclusive, then we use a sigmoid function on each raw output independently. The sigmoid will allow us to have high probability for all our classes, some of them, or none of them. Example: classifying diseases in a chest x-ray image. The image might contain pneumonia, emphysema, and/or cancer, or none of those findings.

If we have a multi-class classification problem that is there is only one "right answer" in which the outputs are mutually exclusive, then we use a soft max function. The soft max will enforce that the sum of the probabilities of our output classes is equal to one, so to increase the probability of a particular class, our model must correspondingly decrease the probability of at least one of the other classes. Example: classifying images from the MNIST data set of handwritten digits. A single picture of a digit has only one identity which is true - the picture cannot be a 7 and an 8 at the same time.

5. Why is logistic regression a type of classification technique and not a regression? Name the function it is derived from?

Logistic Regression is one of the basic and popular algorithms to solve a classification problem. It is named 'Logistic Regression' because its underlying technique is quite the same as Linear Regression. The term "Logistic" is taken from the **Logit function** that is used in this method of classification.

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X.

Contrary to popular belief, logistic regression is a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.
Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself.
If we try to apply Linear Regression on the same problem as a Logistic regression one, we are likely to get continuous values using the hypothesis of linear regression. Also, it does not make sense for h(x_i) to take values larger than 1 or smaller than 0.

Simple Classification with Python

Dataset: https://www.kaggle.com/c/dogs-vs-cats/data

Reference: https://www.geeksforgeeks.org/project-idea-cat-vs-dog-image-classifier-using-cnn-implemented-using-keras/

I have implemented the **Simple Classification** algorithm over the given dataset, and I have run it in Jupyter notebook. The “.ipynb” file and the dataset along with the graph shown is implemented and the same is uploaded on my GitHub repository, the link to which is given below.

GitHub Repository: https://github.com/aiqqia/Machine-Learning

THE END