**ASSIGNMENT 1**

Question - 1:

Explain PCA in details

Answer:

PCA or Principal Component Analysis can be defined in a couple of ways:

1. It can be described as a dimensionality reduction technique which can be helpful while visualising big data systems where there are so many dimensions or features in data that it becomes impossible to plot data so to visualise trends in data.
2. It can also be described as a way in which we can keep the data in the direction of maximum variances such that there is minimal loss of most important information across features, the method to do this (Eigen Value decomposition) ensures the data in the new space or score space is linearly independent

Also, it is important that both these are related as in we cannot achieve dimensionality reduction without going through matrix decomposition anyways. It’s just a point of view from which one might look.

Question – 2:

What is cross validation?

Answer:

We usually split data in training and testing sets (Not done in this assignment since the data set had only 150 data points) when we have enough data points to test our models performance in some sort of way. The problem is that if we do model training only once there is a chance of overfitting on that set of data only and the error/accuracy rating from that model’s performance would probably not show the entire picture.

So, in cross validation what we do is we divide the training data into different folds as well.

Let’s say that we had 1000 training data points and we divide it in 10 folds, we’ll have 100 points in each fold and 10 iterations. Each iteration would have 900 data points and 100 testing points (every other iteration will have different 100 testing points), that would mean 10 errors. After the 10 iterations we get the mean of those errors and that would be a much clearer picture on the performance of the model avoiding overfitting.

We can use grid search as well with cross validation to tune the hyperparameters in these iterations to get the best hyperparameter (E.g.: Optimum number of neighbors on KNN’s model)

Question – 3:

Explain the difference between classification and regression

Answer:

Classification is about predicting discrete values, values that are countable. E.g.: Gender, Hair Color, Petal Color

They are basically of two types:

1. Binary Classification
2. Multi-Class Classification

Regression is about predicting values that are continuous in nature i.e., the values that cannot really be counted discretely. E.g. : Salary, Price

They are basically of two types:

1. Linear Regression
2. Non – Linear Regression

Question – 4:

What is Precision-Recall, F1 score, Confusion Matrix(CM) and Accuracy?

Answer:

Chart

Description automatically generated

Image Source: [Confusion Matrix for Your Multi-Class Machine Learning Model | by Joydwip Mohajon | Towards Data Science](https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826)

Confusion Matrix : Confusion Matrix is a square matrix that shows you the count of:

1. True Negatives : Data Points that were predicted to be Class B and belonged to Class B as well
2. True Positives : Data Points that were predicted to be Class A and belonged to Class A as well
3. False Positives : Data Points that were predicted to be Class A but belonged to Class B in truth
4. False Negatives : Data Points that were predicted to be Class B but belonged to Class A in truth

So, it is a way to concretely show how well our model did by showing the classification correctness data in a matrix or 2D List

Also, the thing to note is that the above matrix is for a Binary classification. If there were 3 classes to be classified, the matrix would have been 3\*3 and so on for each consequent count of True positives and False positives

Precision : Keeping the above under consideration, Precision as a metric tells us that out of the number of Data Points that we identified as belonging to True class, how many were in fact belonging to True Class

Precision = True Positives / (True Positives + False Positives)

Recall : Recall at the same time tells us that out of the number of actual data points belonging to True Class, how many were we able to identify correctly as belonging to True Class

Recall = True Positives / (True Positives + False Negatives)

F1 Score : F1 score is basically a combination of the above two, in that it is the Harmonic mean of precision and Recall. Point to be noted is that all the above values are calculated on the assigned test data only and not all the available data set. So, in a way F1 Score contributes to judgement of accuracy of model on the test data.

F1 Score = 2 \* ((Precision \* Recall) / (Precision + Recall))

Accuracy : Accuracy is the total number of predictions that we got correct out of the total data point predictions that we had to make, it is a high level performance indicator for our model.