**PGP in AI/ML**

**Classification - Assignment 1**

**Submission Date: 2359 hrs on 24-08-2019**

**Total Marks: 25**

The following questions are to enhance your understanding of basic concepts and definitions. You are expected to answer these questions on your own without referring to internet.

1. List the limitations of some prediction measures (like Accuracy). Provide the scenarios highlighting the limitations **[6M**]

**Answer 1:**

There are 3 prediction measures,

1. Accuracy
2. Precision
3. Recall

Accuracy measure have the following limitation,

* Accuracy for a given model could be High. But “False Negative” rate could be still high that would make the model terrible.
* Model with High accuracy could still predict sample as “Positive” while the actuals could be “Negative”. This indicates high “False Positive” rate, that could be bad for the model, but not as bad as “False Negative” cases.
* Accuracy metric works only when the probability of all classes are equally likely possible in the training set.

Below two methods gives biased result.

Precision measure have the following limitation,

* Precision indicate rate of errors due to mis-qualification caused for positive cases only.
* Relying on precision metric could give us a model that could still have high “False Negative” cases.

Recall measure have the following limitation,

* Recall captures only the sensitivity of the model.
* This measure gives us indication only about mis-qualification of False Negative cases. A higher value of Recall cannot guarantee the model is a good model.

When we create the model, we must know what model can tolerate based on the domain of the problem.

For e.g. if the model predicts whether the patient has a chance of “Cancer” or not, then a model having high “False Negative” rate will turn out to be terrible.

However, if the problem domain is to classify house based on the crime rate index of the house location, then higher rate of “False Positives” will classify the property wrongly, and many people will end up buying such property and cursing the model later.

1. Suggest measures countering these shortfalls. **[4M]**

**Answer 2 :**

Having a high accuracy doesn’t give the confidence to deploy the model and expect the model will serve predicting all negative cases as negative, and all positive cases as positive with the accuracy score given.

Measure like precision and recall is used to judge the model for False Positive and False Negative cases. And depending on the business toleration level, we must choose the model with high precision and/or high recall rates.

The shortfall of the above measures, can be somewhat calibrated using few measure below.

* **F1 score,**

This is a weighted average of precision and recall. Keeping a F1 score closer to 1 is of best interest.

*F1\_Score = [ 2 \* (precision \* recall) ] / (precision + recall)*

* **AUC ROC**

In AUC ROC, we consider having a classification threshold that can be calibrated to attain either high recall, high precision and a balance between these measures.

ROC curve is plotted against FPR (False Positive Rate) and TPR (True Positive Rate) which are given as below,

FPR = FP / FP + TN

TPR = TP / TP + FN

And the goal is to achieve the maximum point where the blue curve touches the upper side of the figure, which may be is when FPR = 0.5 and TPR = 0.83.

1. For the dataset -2 (liver disease dataset uploaded along with the assignment), implement kNN classifier using Python. Find out the optimal k for the case study and put up appropriate reasoning/arguments for choosing the optimal k. **[15M**]

**Answer 3:**

**Enclosed Jupiter notebook**

Submission Details

Text answers – id\_classification\_assignment1.doc

Code - id\_knn.ipynb

Contact Details

(You should put up queries in the discussion forum of the corresponding assignment folder only)

1. Himanshu - himanshuchawla437@gmail.com
2. Mohit - mohitrathore15@gmail.com
3. Parthasarathy - pdparthasarathy.03@gmail.com