**Instructions**: Please complete and submit your work to the appropriate folder in LumiNUS. You may work in study groups, but each student must be responsible for their own submission.

Please submit all the following documents as a single zip file named StudentID-Name-HW4.zip:

1. Completed Word file named as StudentID-Name-HW4.docx (with all results)
2. Print preview of ipynb file named as StudentID-Name-HW4.pdf (with all results)
3. Working ipynb file named as StudentID-Name-HW4.ipynb
4. Consider the metrics accuracy, precision, and recall.
   1. Give one example when accuracy would not be a good performance metric. Give a numerical example.

**Answer:**

Assume there is an imbalance between positive and negative samples, 990 were positive and 10 were negative. A model that predicts every sample is positive, and its accuracy rate is 99%. It seems that the accuracy is high, but this model is not a good model.

* 1. Given one example of a supervised machine learning classification problem when higher precision is desired. Please give a different example than the ones given in class. This need not be a numerical example but must be clearly defined classification problem and dataset.

**Answer:**

Let's assuming we're using our phone's face recognition to unlock, and our dataset is a bunch of photos, positive for the phone's owner photo, negative for someone else. So the precision must be very high, or a stranger will be able to unlock the owner's phone

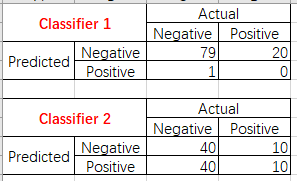
* 1. Given one example of a supervised machine learning classification problem when higher recall is desired. Please give a different example than the ones given in class. This need not be a numerical example but must be clearly defined classification problem and dataset.

**Answer:**

Suppose we are testing for COVID-19. Our dataset is samples of many people. Positive means the person is a COVID-19 patient, and negative means he is not. Recall rates are really important, otherwise there will be people who are patient but not detected and spread COVID-19.

1. Suppose you are given the same test dataset and two binary classifiers. Give a numerical example such that Classifier 1 has higher accuracy than Classifier 2, but Classifier 2 has both higher precision and higher recall than Classifier 1? Hint: Give a hypothetical 2x2 confusion matrix for each classifier.

**Answer:**



The accuracy of Classifier 1 is 79%, and accuracy of Classifier 2 is 50%.

Precision and recall of Classifier 1 is both 0%, Classifier 2 is 20% and 50%

1. In this problem, we will look at the Breast Cancer Wisconsin (Diagnostic) Data Set available UCI Machine Learning Repository. Please use the wdbc.data dataset from:

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

* Use this dataset to predict cancer diagnosis (malignant or benign). Use a random train/test data split of 70%/30%. Repeat this process 20 times and compute the average performance.
  1. Please evaluate the following algorithms: Random Forest, SVM with RBF kernel, and Multilayer perceptron. Please compute the following metrics and fill in the table below. You can try various parameters for these algorithms and report your best results.
  2. Please also plot the ROC and PRC curves for both classifiers. Please also compute the AUROC and AUPRC.

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_curve.html>

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | | Precision | Recall | AUROC | AUPRC |
|  | Training Set | Test Set |  |  |  |  |
| Random Forest | 0.9422 | 0.9006 | 0.9167 | 0.7719 | 0.9831 | 0.97 |
| SVM with RBF kernel | 0.9698 | 0.9474 | 0.9286 | 0.9123 | 0.9862 | 0.98 |
| MLP | 0.9497 | 0.9357 | 0.8966 | 0.9123 | 0.9826 | 0.97 |