**BUSINESS DATA MINING**

**(IDS 572)**

**Solutions to Homework 4**

**Group Members**

* Amey Pophali (apopha2@uic.edu)
* Karthik Varanasi (vvaran3@uic.edu)
* Mrinal Dhawan (mdhawa3@uic.edu)

**Problem 1 –**

**Take a look at three examples below and answer the following questions:**

****

* **What can you say about this ROC curve? How this classifier differs from a random guess? Pick**

**one point on a curve and interpret it using examples and illustrations. For example, this point**

**represents a classifier that can detect x% of all patients, who have a disease, but y% those who**

**have not, are classified incorrectly....**

Answer -

This is an average ROC curve which shows the sensitivity vs specificity. This curve describes an algorithm that performs better than a random guess as the area under the curve is 0.835 which is greater than 0.5 which is the area under the curve of a random guess.

****

* **(b) Compare two ROC curves above. Which one is a better model and why?**

Answer -

Between Algorithm A and Algorithm B, Algorithm A is better because even though Algorithm B has more sensitivity in the final data, Algorithm A is consistent in the results.

* **(c) Compare two ROC curves above. When algorithm A would be preferred over algorithm B?**

Algorithm A is preferred over algorithm B as long as its ROC curve is above the Algorithm B’s and from the point where the Algorithm B’s curve moves above the Algorithm A, we prefer Algorithm B over Algorithm A because it has a better chance of classifying the data when compared to algorithm B.

****

**Problem 2 –**

**Calculate (on paper) confusion matrix, precision and recall for the given dataset under**

**threshold of 0.5.**

**Draw a ROC curve.**

****

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted | |
|  |  | Negative | Positive |
| Actual | Negative | 3(TN) | 2(FP) |
| Positive | 0(FN) | 5(TP) |
|  |  |  |  |
|  |  |  |  |
| Precision | 0.714286 | =TP/(TP+FP) |  |
|  |  |  |  |
|  |  |  |  |
| Recall | 1 | = TP/(TP+FN) |  |

|  |  |
| --- | --- |
| Sensitivity | 1 |
| =TP/TP+FN |  |
| Specificity | 0.6 |
| =TN/FP+TN |  |
| 1-speicifity | 0.4 |

**Problem 3 –**

**Assume that two individuals offer to sell you their predictive models M1 and M2. The**

**confusion matrices produced by each model are as follows.**

**(a) Assuming that precision is of paramount importance in your application, which of the two**

**models would you buy? Why?**

****

**(b) Assuming that the cost of labeling as True something that is actually False far exceeds the cost**

**of labeling as False something that is actually True, which of the two models would you buy?**

**Why?**

****