# CS909 Lab 7: Data Mining

February 27, 2014

**Student Name:**

**University ID:**

**7.1 Suppose that you have used some concept learning algorithm to learn a hypothesis h1 from some training data. You are interested in knowing the accuracy that the hypothesis can be expected to achieve on the underlying population. You assess the hypothesis on a set of test data consisting of 45 instances and you observe an error of 6.67%. Calculate the 95% interval for the expected error.**

| **Mark:** | *None —— Basic —— Good —— Excellent* |
|---|---|
| **Comments:** | |

**7.2 You compare two supervised learning classification algorithms on a set of training data and you find out that 10-fold cross validation yields the following accuracies.**
**At what confidence level can you assume that algorithm 1 will outperform algorithm 2? Show your working.**

| **Mark:** | *None —— Basic —— Good —— Excellent* |
|---|---|
| **Comments:** | |

**7.3** **Ravensworth, Liakata and Clare trained a classifier on predicting the type of a scientific article (e.g. Review, Research etc.) according to PlosOne categories and using as features the distribution of Core Scientific concepts within papers (e.g. Hypothesis, Methods, Conclusion etc.)**
**Perform the same classification task on the data set AI2013_papers.csv using random forests, naiveBayes and SVM from the e1071 R package.**
**Compare classifier performance using 10-fold cross-validation and report appropriate measures. Give confidence levels for the accuracy of the best performing classifier.**
**If you get a higher F-score on some of the classes please let us know!**

| **Mark:** | *None* —— *Basic* —— *Good* —— *Excellent* |
|---|---|
| **Comments:** | |

**I confirm that the work that has been marked is my own and understand that cases of plagiarism will be subject to Departmental and University regulations.**

**Student Signature** **Marker Signature**