# CS909: 2017-2018 (MSc Only)

**Exercise two: Perceptron, SVM, ensemble methods, classifier evaluation, clustering, text classification**

**Submission: 12 pm (midday) Wednesday 25th April 2018**

**Notes:**

**(a) The use of libraries is not allowed as part of this exercise, unless explicitly permitted.**

**(b) Your submission should consist of both code and a report explaining what you have done.**

**(c) This assignment will contribute to 35% of your overall mark.**

## Part A: Perceptron (30)

**1.** Implement a perceptron classifier for linear binary classification using R. The algorithm presented on slide 14 (Algorithm 7.1 in the Flach book) should be sufficient for your implementation. Create a dataset that only includes the versicolor and setosa species of iris. Use your binary perceptron to classify this new iris dataset. What if the data is presented in random order? What do you observe? Experiment with different learning rates η and a different value for the intercept b. Would your algorithm always converge? Report and comment on all of your experiments .

## Part B: classifier evaluation, SVM, ensemble methods, clustering (55)

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 145 instances and you observe an error of 6.67%. Manually calculate the 95% interval for the expected error. Include your working in the submission. (10)

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:

| CV Fold | Algorithm 1 | Algorithm 2 |
|---|---|---|
| 1 | 91.11 | 90.7 |
| 2 | 90.48 | 90.52 |
| 3 | 91.87 | 90.88 |
| 4 | 90.52 | 90.87 |
| 5 | 89.88 | 90.02 |
| 6 | 89.77 | 88.99 |
| 7 | 91.44 | 90.98 |
| 8 | 90.88 | 91.44 |
| 9 | 90.77 | 90.77 |
| 10 | 90.89 | 90.92 |

At what confidence level can you assume that algorithm 1 will outperform algorithm 2? Show your working in the report. (10)

3. Ravensworth, Liakata and Clare [1] trained a classifier on predicting the type of a scientific article (e.g. Review, Research etc.) according to PlosOne categories ("type") and using as features the distribution of Core Scientific concepts within papers (e.g. Hypothesis, Methods, Conclusion etc.).

   **a.** Ignore the target value "type" and use three clustering algorithms of your choice, selected from those discussed in lectures. You can select these from suitable R packages. Justify your choice of clustering algorithm. Provide and implement appropriate measures of cluster quality. Is there a correspondence between clusters and the original type labels? (35)

   Include all your working and code in your submission and clearly report what you have done and why.

   [1] J. Ravenscroft, M. Liakata, A. Clare (2013). Partridge: An effective system for the automatic classification of the types of academic papers. In Proceedings of AI-2013.

## Part C: Text classification (MSc ONLY) (110)

### Objective

The objective of this exercise is to evaluate your understanding of representing documents as a set of features and performing classification on the documents, as discussed in lectures.

To do this exercise you will need to use the reuters.csv file from the website.
This data set is one of the standard corpora used by text mining researchers to test their algorithm and in its original format it consists of 21 SGML files, eachwith multiple tags. To facilitate the task we present the files as one large .csv where each row represents a document and columns indicate which topics are

relevant to a document, the title of the document and the actual text in the document. You can use either Python or R for this exercise as well as corresponding libraries. It is important to explain clearly what you have done and why and include appropriate results and table matrices

**Tasks**

1. Explore the data and undertake any cleaning/pre-processing that you deem necessary for the data to be analysed. Focus your efforts on the 10 most populous classes, namely: (*earn, acquisitions, money-fx, grain, crude, trade, interest, ship, wheat, corn*), so only consider documents belonging to these topics.

2. Obtain ngram features of the documents/news articles as discussed in the text mining lectures. Your final report must provide a summary of your assumptions regarding feature representation as well as an explanation of how you obtained the features.

3. Build classifiers, using either R or Python libraries to predict the TOPICS tags for documents. Use the training data for building the models and comparing their accuracy, precision and recall. You can reuse your code from Part B, exercise 3c. Explain whether you are using micro or macro averaging and why. Use the test set only to get an estimate of the accuracy, precision and recall of the "optimal" model based on your analysis using the training data

You will be evaluated on the basis of the report, you will not have the chance to explain your work in person, so this has to be well written and clear. Include:

a. A description of the data, any pre-processing performed on it and why. (30)
b. Explanation of feature representation and details on any feature selection implemented. (30)
c. Details on which classification algorithms were used and why. What parameters were used in each case and why. (20)
d. Perform full evaluation of the classification algorithms. Compare classifier performance using 10-fold cross-validation (which you should implement yourself) and report appropriate measures (you should also implement the measures). Give confidence levels for the accuracy of the best performing classifier. (30)

Include your code in the submission. Where appropriate in the report provide references to your code. Some of the code can go in the appendix.