MPhil in Advanced Computer Science
Machine Learning for Language Processing

Leader:	Prof Ted Briscoe and Dr Mark Gales
Timing:	Lent term
Prerequisites:	None, Intro to NLP and Stat. Sp Proc & Lg Modelling useful
Structure:	8 Lectures and 8 Seminars

AIMS

This module aims to provide an introduction to machine learning with specific application to tasks such as document topic classification, spam email filtering, and named entity and event recognition for textual information extraction. We will cover supervised, weakly-supervised and unsupervised approaches using generative and discriminative classifiers based on graphical models, including hidden Markov models, Gaussian mixture models and conditional random fields.

SYLLABUS

1. Classification by machine learning – classification vs. prediction, types of classifier, generative vs. discriminative models, supervised training (2Ls, Dr Gales)

2. Document Topic Classification – bag-of-words representation, evaluation measures, feature selection, model comparison (2Ss, Prof Briscoe)

3. Graphical Models – Markov Models, Hidden Markov Models, Gaussian Mixture Models, Conditional Random Fields, Estimation Maximisation, Variational Inference (4Ls, Dr Gales)

4. Spam Email Filtering – task, adaptive training, semi-structured documents, N-gram language models, evaluation (1S, Prof Briscoe)

5. Named Entity Recognition – HMMs vs. CRFs vs. Parsing & non-sequential classification, inherent vs. contextual features, feature dependence, partial vs. actively labelled data, evaluation (2Ss, Prof Briscoe)

6. Support Vector Machines – maximum margin classifiers, kernel "trick", types of kernel (1L, Dr Gales)

7. Relation Extraction – sequence vs. tree vs. graph kernel approaches, evaluation (2Ss, Prof Briscoe)

8. Clustering – factor analysis, singular value decomposition, principal component analysis, (1L, Dr Gales)

9. Document Topic Clustering and Term Clustering – latent semantic indexing / analysis, incremental semantic analysis, evaluation (1S, Prof Briscoe)

OBJECTIVES

On completion of this module students should:

- understand the issues involved in applying machine learning approaches to a range of language processing applications;
- understand the theory underlying a number of machine learning approaches that have been applied to language processing, including: graphical models, Gaussian mixture models, conditional random fields, support vector machines;
- understand applications of machine learning to specific tasks including: document topic classification and clustering, SPAM filtering, named entity recognition.

COURSEWORK

Students will be expected to undertake reading for assigned lectures and seminars. Each student will give a 20-30 minute presentation of one paper during a seminar.

PRACTICAL WORK

None

ASSESSMENT

- Students will receive one tick worth 5% for attendance at seminar sessions, reading of assigned material, and satisfactory contribution during seminars.
- Students will receive a second tick worth 5% for a satisfactory presentation of an assigned paper.
- Students will write an in-depth essay on a topic agreed with the lecturers of not more than 5000 words. The essay will be due at the beginning of the Easter Term, will be assessed by one of the lecturers, and will account for 90% of the module marks.

RECOMMENDED READING

- Christopher Bishop: *Pattern Recognition and Machine Learning*, Springer, 2006. (Chaps: 1,2, 7-10, 12)
- Daniel Jurafsky and James Martin *Speech and Language Processing*, 2nd edition, Prentice-Hall 2008 (Chaps: 4,6,22)
- Christopher Manning, Prabhakar Raghavan and Hinrich Schutze, *Introduction to Information Retrieval*, Cambridge University Press. 2008 (Chaps: 12–18)

Last updated: May 2010