UNSW SYDNEY

FACULTY OF SCIENCE
SCHOOL OF MATHEMATICS AND STATISTICS

MATH5836

Data Mining and its Business Applications

Semester 2, 2017

MATH5836 – Course Outline

Information about the course

Course Authority and Lecturer: Dr. Zdravko Botev, email `botev@unsw.edu.au`

Credit: This course counts for 6 Units of Credit (6UOC).

Prerequisites: It is expected that students in this course have a solid understanding of three to four years worth of probability and statistics at an undergraduate level. In addition, it will be assumed that students have had first experiences with statistical software packages such as R or Matlab.

Tutorials/Labs: There will be no formal tutorial/labs for this course. There will be assignments which will require students to undertake computational laboratory work. However, depending on student demand for consultation and advice there may be tutorials in some of the weeks prior to the due date of an assignment. We have booked the Red Center RC-4082 on all Tuesdays from 15:00 to 16:00 in case there is a need for such consultation/tutorial.

Lectures: Mondays, 5:00pm -8:00pm in Room LG03 (The Michael Hintze Theatre in the Tyree Building, or Tyree TETB LG03). This is box H6 on the campus map. The Michael Hintze theatre can be accessed from near Consultation room LG32, or the entrance near the elevators that also connects to the Horse shoe theatre.

Course aims

This course will aim to provide the student with a working knowledge of a select topics from data mining and machine learning. Particular focus will be on the fundamental statistical properties and analysis of a select few popular techniques for learning, classification and prediction. This course will mostly include a hands-on experience with implementing algorithms and running code to solve computationally intensive statistical problems. The course will also include a limited development of solid theoretical (theorems follows by proofs) analysis of classification error in the simplest machine learning settings.

Relation to other statistics courses

This course complements Multivariate Analysis - MATH5855; Statistical Computations - MATH5856. It is similar in content but different in approach to Machine Learning and Data Mining - COMP9417.

Student Learning Outcomes

This course is expected to give students an understanding of the fundamentals of machine learning and basics of data mining, which is essential for anyone contemplating a career as a professional statistician or data analyst in industries reliant upon such expertise. The student should develop a working knowledge of the statistical and theoretical underpinnings of the topics covered. Given this fundamental statistical understanding of these methodologies this will allow the student to utilise these techniques with confidence on real world data sets and scenarios. As such the student is expected to develop applied working knowledge of the methodologies covered, largely through application in assignments during semester. It is stressed that this course is aimed at fundamental statistical properties of these methods, it is not a course on application of computer software.

Relation to graduate attributes

The problem-solving activities in assignments will improve your research, inquiry and analytical thinking abilities (Science Graduate Attribute 1) and your capacity and motivation for intellectual development (Science Graduate Attribute 2); Coursework assignments will provide you with timely feedback on your progress and improve your Communication skills (Science Graduate Attribute 4); Computing skills developed in this course will improve your Information Literacy (Science Graduate Attribute 6)

Teaching strategies underpinning the course

New ideas and skills are introduced and demonstrated in lectures and through recommended reading of supplementary material such as research papers, then students develop these skills by applying them to specific tasks in assessments.

Rationale for learning and teaching strategies

We believe that effective learning is best supported by a climate of inquiry, in which students are actively engaged in the learning process. Hence this course is structured with a strong emphasis on problem-solving tasks. Students are expected to devote the majority of their class and study time to the solving of such tasks. New ideas and skills are first introduced and demonstrated in lectures, and then students develop these skills by applying them to specific tasks in assessments. Computing skills are developed and practiced in assignment working groups. This course has a major focus on research, inquiry and analytical thinking as well as information literacy. We will also explore capacity and motivation for intellectual development through the solution of both simple and complex mathematical models of problems arising in the quantitative sciences and the interpretation and communication of the results.

Assessment

Assessment in this course will consist of a final exam, two group assignments, and class/tutorial/lab participation. See table below.

Assessment	Weight	Date
Final Exam	60%	
Assignment 1	20%	Week 6
Assignment 2	15%	Week 12
Class Participation	5%	Throughout semester
total:	100%	

Knowledge and abilities assessed: All assessment tasks will assess the learning outcomes outlined above, specifically, the ability to derive logical and coherent proofs of relevant results, and the ability to solve a variety of problems, both theoretical and in practice. In addition, through group assignments the students form practical time and relationship management skills. A single report will be required per group.

Assessment criteria: The main criteria for marking all assessment tasks will be clear and logical presentation of correct solutions.

Final Exam

Rationale: The final examination will assess student mastery of the material covered in the lectures and assignments.

Assignments

Rationale: There will be two assignments for students to try their hand at more difficult problems requiring more than one line of argument or computation. These assignments will be assigned to groups. Late assignments will not be accepted. Note UNSW's policy on plagiarism: `www.lc.unsw.edu.au/plagiarism`

Class/Tutorial/Lab Participation

Rationale: Tutorial/laboratory problems will be given, mostly taken from the textbooks. Students will be given the opportunity to present solutions to tutorial/laboratory problems during the tutorials/lectures/laboratories. The lecturer will keep track of each students' participation on a scroll.

Additional resources and support

Consultation and Moodle discussion forums

Students are encouraged to make extensive use of the discussion forum on moodle. In addition, students are encouraged to ask questions regarding the material during consultation hours/tutorials/labs.

Please email the lecturer if an issue requires confidentiality and urgency (e.g., illness, bereavement, etc.). In that case, the lecturer will make individual appointments to resolve the issue.

Computer laboratories

Computer laboratories (RC-M020 and RC-G012) are open 9:00am – 5:00pm Monday – Friday on teaching days. RC-M020 has extended teaching hours (usually 8:30am – 9:00pm Monday – Friday, and 9:00am – 5:00pm Monday-Friday on non-teaching weeks).

Lecture notes

All lecture slides will be provided on moodle before or after each respective lecture. These notes are *insufficient* to understand the course material, therefore there will be provided detailed repositories of additional recommended texts. This is expected to also be read to supplement the material covered in lectures.

Course Evaluation and Development

The School of Mathematics and Statistics evaluates each course each time it is run. We carefully consider the student responses and their implications for course development. It is common practice to discuss informally with students how the course and their mastery of it are progressing.

Administrative matters

School Rules and Regulations: Fuller details of the general rules regarding attendance, release of marks, special consideration etc. are available at are available at `http://www.maths.unsw.edu.au/currentstudents/assessment-policies`. For UNSW policies, procedures and guidelines see `https://student.unsw.edu.au/policy`

Plagiarism and academic honesty: Plagiarism is the presentation of the thoughts or work of another as ones own. Issues you must be aware of regarding plagiarism and the universitys policies on academic honesty and plagiarism can be found at `https://student.unsw.edu.au/plagiarism`

Tentative Course Schedule

Week 01 Linear models for regression; Shrinkage methods
Week 02 Linear Models for classification
Week 03 Theoretical foundations: Probability inequalities and concentration results
Week 04 Vapnik-Chervovenkis (VC) theory and empirical risk minimization
Week 05 Applications of VC theory to linear classification
Week 06 Kernel Smoothing Methods and Splines
Week 07 Projection Pursuit; Neural Networks
Week 08 Unsupervised Learning: Mixture models; Clustering
Week 09 Latent variables; Expectation Maximization Algorithm
Mid-Semester Break
Week 10 Bayesian inference; Markov chain Monte Carlo; Sequential Monte Carlo
Week 11 Classification and Regression Trees; Multivariate Adaptive Regression Splines
Week 12 Revision or Bagging and Boosting (depending on time)