KXT711 Data Mining and Text Retrieval

Semester 1, 2010
Sandy Bay Campus, Hobart
Newnham Campus, Launceston

Unit Coordinator
Dr. Mike Cameron-Jones
E-Mail: Michael.CameronJones@utas.edu.au
Phone: (03) 6324 3395
Room: V171, Newnham Campus, Launceston

Lecturing Staff
Newnham Campus, Launceston:
Dr. Mike Cameron-Jones
E-Mail: Michael.CameronJones@utas.edu.au
Phone: (03) 6324 3395
Room: V171

UNIT OVERVIEW

Introduction
Machine Learning is commonly considered to be a sub-field of Artificial Intelligence, and can be seen as the study of computational approaches to finding patterns in data. Data Mining applies Machine Learning techniques to look for patterns in large data sets. Many of the applications of Data Mining involve textual data, such as e-mails (e.g. when spam filtering) or Web pages (e.g. when searching for information).

This unit introduces the key ideas and techniques in Machine Learning in general, and some techniques for working with textual data in particular, including techniques from the field of Information Retrieval.

Prerequisites
(KXT201 or KXA251) & (KXT206 or KXT306 or KXA252)

Unit Weight
12.5% of one academic year

Teaching Pattern
Lectures: 3 hr/wk
Tutorials: 1 hr/wk (from week 2)

Unit Content
Classifier Learning:
1. Decision Trees
2. Rule Sets
3. Naive Bayes
4. Instance Based Learning
5. Neural Networks
6. Other and Combined Techniques
Other Data Mining / Machine Learning Topics:
1. WEKA
2. Numeric Prediction
3. Techniques for Other Tasks, e.g. Clustering

Text Retrieval:
1. Traditional Information Retrieval
2. Structured, and / or Linked, Text
3. Machine Learning in Text Retrieval

For more information see the section titled ‘Content’ on the unit website.

**Prior Knowledge and/or Skills**

Students entering the unit will be assumed to have a knowledge of algorithms, of Artificial Intelligence, and of programming in Java, as per the unit's direct and indirect prerequisites.

**Learning Outcomes**

On successful completion of this unit, you will be able to:

1. Demonstrate that you understand the key ideas and techniques in machine learning / data mining sufficiently to apply them.
2. Demonstrate that you understand the key ideas and techniques in machine learning / data mining sufficiently to participate in masters by coursework level research in the area.
3. Demonstrate that you understand the basics of text retrieval techniques.

**Generic graduate attributes**

The university has defined a set of generic graduate attributes expected in its graduates. 

Your course is designed to enable you to develop generic skills that are valued in, and expected of, graduates. These are skills that you will need to develop over time. Hence you are encouraged to look for opportunities, as you study each unit, to reflect on and improve these skills.

In this unit these skills are specifically targeted:

**Knowledge:** Students will have the opportunity to increase their knowledge of the areas of machine learning / data mining and text retrieval — techniques that are increasingly widely used.

**Communication skills:** Students will have the opportunity to practise their written communication skills.

**Problem-solving skills:** Students will have the opportunity to work with others.

**UNIT ASSESSMENT**

**Assessment Pattern**

In-semester (30%), exam (70%)

**Assessment Summary**

<table>
<thead>
<tr>
<th>Component</th>
<th>Weight</th>
<th>Due Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work before, and in, assessable tutorials</td>
<td>10%</td>
<td>During semester</td>
</tr>
<tr>
<td>Assignment</td>
<td>20%</td>
<td>3pm 27th May (Thursday of Week 13)</td>
</tr>
<tr>
<td>3 hr Examination</td>
<td>70%</td>
<td>University Examination Period</td>
</tr>
</tbody>
</table>

**Assessment Items**

**Item 1**

**Title:** Work before, and in, assessable tutorials  
**Type:** In-Semester - learning tasks  
**Task Length:** not applicable  
**Weighting:** 10%  
**Links to Learning Outcomes:** 1, 3  
**Due:** During semester  
**Description:** During the semester students will be expected to prepare for and participate in their allocated tutorials. In 7 weeks of the semester this will be an assessable matter. (Which 7 weeks will be specified during the semester.)

Satisfactory preparation for, and satisfactory participation in, at least 5 of a student's assessable allocated tutorials will be worth 10%. Satisfactory preparation for, and satisfactory participation in, at least 4 (but not 5) of a student's assessable allocated tutorials will be worth 5%. No marks will be awarded for fewer than 4.

Late submissions are not accepted for this assessment item. The requirement for full marks has been set at 5 of the 7 assessable tutorials to allow a margin for such things as family or work commitments. Where an assessable tutorial does not occur, e.g. due to a public holiday, students allocated to that
tutorial may count it towards their (4 or) 5. Where a student misses an assessable tutorial due to appropriately certificated illness, that student may apply to the unit coordinator to count that tutorial towards their (4 or) 5.

**Item 2**

**Title:** Assignment  
**Type:** In-Semester - group assignment  
**Task Length:** not applicable  
**Weighting:** 20%  
**Links to Learning Outcomes:** 1, 2  
**Due:** 3pm 27th May (Thursday of Week 13)  
**Description:** The assignment will involve implementing or modifying some machine learning programs, experimenting with them and others, and writing a small paper reporting the results obtained. Groups will be of size at most 2.

**Item 3**

**Title:** 3 hr Examination  
**Type:** Formal Examination  
**Task Length:** 3 hours  
**Weighting:** 70%  
**Links to Learning Outcomes:** 1, 2, 3  
**Due:** University Examination Period  
**Description:** It is expected that this year's exam will be broadly similar to last year’s KXT311 exam in respect of the style of questions, assessment balance amongst major topics, and materials permitted. Further information regarding this year's exam will be given towards the end of the unit.

See the 'Assessment' section in unit website for more detailed information about assessment items.

**How your Final Grade will be determined**

Overall assessment will be based on the student’s performance throughout the semester as well as in a formal examination. In order to achieve a pass (or better) result, a student must obtain:

1. at least 45% of the total mark for in-semester assessment items  
2. at least 45% of the mark for the formal examination  
3. at least 50% of the overall mark

**UNIT RESOURCES**

**Unit Web Site**

This unit is Web Dependent: content. This means that you will need to use the Web for this unit. The unit website contains unit information and resources. The unit website is accessed from http://www.utas.edu.au/coursesonline/. You will need to use your university email pop account username and password to log on to the MyLO system. Once authenticated by the system your personalised MyLO Learning Online area will be displayed. It contains links to the websites that you have permission to access - including the website for this unit. If you are not able to access the unit website, please contact the University IT help desk:

- Entrance Level, Morris Miller Library, Sandy Bay Campus;  
- Entrance Level, Launceston Campus Library, Newnham Campus.  
- Telephone: 6226 1818 and 1300 304 903.  
- The 1300 number is a local call from within Tas, with the exception of mobiles.  
- Email: servicedesk@utas.edu.au  

**Prescribed Text**


**Software**

The software that you will need to access the unit website and to study this unit, including general purpose software such as word processors, is provided on the computers in the School's computing labs. If you intend to use software on other computers please check that the versions are compatible. The unit will use the WEKA suite of machine learning programs (in Java).

**GENERAL RESOURCES**

**School Website**

School of Computing and Information Systems - Faculty of Science, Engineering, and Technology.  

**Faculty Website**
Information and Resources for Faculty of Science, Engineering and Technology students are available on the faculty website at: http://www.utas.edu.au/scieng

University Website

Information and Resources for 'Current Students' are available on the university website at: http://www.utas.edu.au/students/

School Help Desk

Contact the School Help Desk if you have any queries or problems with accessing, using, or printing from the computers in the School of Computing and Information Systems labs.

In Hobart the Help Desk is located on level 3 in the Centenary Building, and is open from 10:00am-12:00pm, and 2:00pm-4:00pm Monday-Friday. The phone number is 6226 2929.

In Launceston the Help Desk is located near the entrance to the computing labs and is open from 10:00am-12:00pm, and 2:00pm-4:00pm Monday-Friday. The phone number is 6324 3447.

Both help desks will accept queries over the phone outside the standard opening hours.

The computer labs at the Cradle Coast Campus are maintained by ITR - please contact the University Help Desk for assistance with these computers.

Computing Facilities

The School has PC labs (running Windows XP), Mac labs (running Mac OS X 10.5), and special purpose Networking labs at the Newnham and Sandy Bay campuses. All students are provided with logins for Windows, Macintosh and Unix environments. If you have not used these facilities before please contact the School Help Desk to collect your account details. If you would like to access these facilities after hours please contact the School Help Desk.

In Hobart, there will be 4 PC Labs, 2 Mac Labs, and 1 Networks Lab in the Centenary Building. In Launceston, there are 2 PC Labs, 1 Mac Lab, 1 Networks Lab, and one Multipurpose Lab in Building V.

Use of Facilities

Use of computing facilities provided by the School is subject to the School's Ethics Guidelines, details of which are posted at http://www.cis.utas.edu.au/cisview/ethics.jsp. Copies of the guidelines are also available in all School labs. The School's facilities may only be used for study-related purposes, and may not be used for personal gain. Anti-social behaviour in labs such as game playing, viewing pornography, loud discussion, audio without the use of head-phones, etc is strictly prohibited in all labs at all times. Eating, drinking, and smoking is not permitted in the labs. Before being granted access to the School's facilities, you will be required to sign a declaration that you have read and understand these guidelines, and that you will abide by them. Disciplinary action may be taken against students who violate the guidelines.

Learning Strategies

If you need assistance in preparing for study please refer to your tutor or lecturer. For additional information refer to the Learning Development website: http://www.utas.edu.au/learndev/

If you will be using MyLO for the first time and would like some information on how to use MyLO refer to the following website: http://www.utas.edu.au/coursesonline/mylo-support.htm

Some of the units you will study use videoconferencing to deliver lectures and tutorials. To enable you to get the best out of a videoconference please refer to the following guide: http://www.its.utas.edu.au/videoconf/vcstudentguide.pdf

Help resolving concerns about this unit

In the first instance you should contact your lecturer. If the matter is not resolved then you should contact the Head of School. If the matter is still unresolved and you would like to know who to contact or the procedures for resolving your concern refer to the following website: http://acserv.admin.utas.edu.au/complaints_info.html

The Hobart based Tasmanian University Union (TUU) or the Launceston/Burnie based Student Association (SA) may also be able to assist.

The School reserves the right to alter the details contained in this Unit Outline. Students will be advised of changes to the outline via their University email account and it remains the responsibility of the student to check their email for such changes.

Occupational Health and Safety

The university is committed to providing a safe and secure teaching and learning environment. For more information see http://www.admin.utas.edu.au/hr/ohs/pol_proc/

University Services and Support

If you are experiencing difficulties with your studies or assignments, have personal or life planning issues, disability or illness which may affect your course of study, you are advised to raise these with your lecturer in the first instance.
The University has staff available to assist you, such as the:

- Learning Development Advisor
- Student Counselor
- Careers Advisor
- Disability Officer

For more information and contact details see the Services and Support section on the University 'Current Students' web page: http://www.utas.edu.au/students/

GENERAL ASSESSMENT

Approach to Learning

The University is committed to high standards of professional conduct in all activities, and holds its commitment and responsibilities to its students as being of paramount importance. Likewise, it holds expectations about the responsibilities students have as they pursue their studies within the special environment the University offers.

The University’s Code of Conduct for Teaching and Learning states:

Students are expected to participate actively and positively in the teaching/learning environment. They must attend classes when and as required, strive to maintain steady progress within the subject or unit framework, comply with workload expectations, and submit required work on time.

You are expected to spend about 130 hrs studying in this unit - this includes attendance at scheduled teaching sessions. (For a 13 week semester this is, on average, 10 hr/wk.) This is the amount of study time that the 'typical' student will need to reach the level of competence and understanding required to fulfil the unit objectives. You are expected to:

- attend all scheduled teaching sessions, unless otherwise notified by the unit coordinator
- prepare for, and actively participate in all scheduled teaching sessions
- complete the assigned learning tasks
- review what has been learnt
- complete assessment items and submit them on time
- access and be familiar with the information and resources available on the unit website
- seek help from teaching staff if you have any questions or difficulties in studying this unit

You are encouraged to read the university's Code of Conduct for Teaching and Learning. Part A describes the 'Responsibility of the University to Students' and part B describes the 'Responsibilities of Students to the University'. http://www.utas.edu.au/policy/code_conduct.pdf

It is expected that students will familiarise themselves with access and use of the MyLO system operated by the University for the electronic delivery of course materials, and for various forms of communication.

It is expected that students will consult email sent to their University email address at least twice a week for notices relating to the administration of the unit, and for notification of the results of assignments.

It is expected that students will read the background material specified in the course curriculum, will actively attend and participate in tutorials, and be prepared to discuss relevant issues arising with tutors, lecturers and fellow students.

Student Expectations of the Unit

Students enrolled in this Unit may reasonably expect the following:

1. To be able to contact a lecturer or tutor by electronic mail, to raise issues arising in the unit, either relating to content or student performance within the unit.
2. Subject to availability, to be able to discuss such issues in person with the lecturer or tutor.
3. That assignments will be marked and the marks will normally be returned within 3 weeks of due dates.
4. That all relevant notices regarding the administration of the unit, including any necessary changes, will be communicated to all students enrolled in the unit via email.

These expectations are in addition to those specified in relevant University regulations.

Plagiarism

Unless specifically stated in the specification of the assessment item provided on the unit website, it is required that:

- work submitted by a student is the work of that student alone OR
- where the assessment item is to be completed by a group of students, the work submitted by the group of students is the work of that group of students alone.

While students are encouraged to discuss the assignments in this unit and to engage in active learning from each
other, it is important that they are also aware of the University’s policy on plagiarism. Plagiarism is taking and using someone else’s thoughts, writings or inventions and representing them as your own; for example downloading an essay wholly or in part from the internet, copying another student’s work or using an author’s words or ideas without citing the source.

"Plagiarism is a form of cheating. It is taking and using someone else’s thoughts, writings or inventions and representing them as your own; for example, using an author’s words without putting them in quotation marks and citing the source, using an author’s ideas without proper acknowledgment and citation, copying another student's work.

If you have any doubts about how to refer to the work of others in your assignments, please consult your lecturer or tutor for relevant referencing guidelines, and the academic integrity resources on the web at http://www.academicintegrity.utas.edu.au.

The intentional copying of someone else’s work as one’s own is a serious offence punishable by penalties that may range from a fine or deduction/cancellation of marks and, in the most serious of cases, to exclusion from a unit, a course or the University. Details of penalties that can be imposed are available in the Ordinance of Student Discipline – Part 3 Academic Misconduct, see http://www.utas.edu.au/universitycouncil/legislation/.

The University and any persons authorised by the University may submit your assessable works to a plagiarism checking service, to obtain a report on possible instances of plagiarism. Assessable works may also be included in a reference database. It is a condition of this arrangement that the original author’s permission is required before a work within the database can be viewed."

It is important that you understand this statement on plagiarism. Should you require clarification please see your unit coordinator or lecturer. Useful resources on academic integrity, including what it is and how to maintain it, are also available at: http://www.academicintegrity.utas.edu.au

Referencing

The preferred text referencing systems for the School is the Harvard system (also referred to as the author-date system). In your written work you will need to support your ideas by referring to scholarly literature, works of art and/or inventions. For information on presentation of assignments, including referencing styles: http://utas.libguides.com/referencing

It is important that you understand how to correctly refer to the work of others and maintain academic integrity. Failure to appropriately acknowledge the ideas of others constitutes academic dishonesty (plagiarism), a matter considered by the University of Tasmania as a serious offence. The university document on plagiarism contains information about referencing the work or ideas of others (see http://www.utas.edu.au/plagiarism/).

Submissions

The details of the submission method (paper, electronic or other) for each assignment will be supplied in a separate assignment specification sheet. All in-semester assignment submissions (including electronic submissions) are to include an Assignment Cover Sheet which includes a statement confirming that the submission is your own work. The Assignment Cover Sheet is available from the School Help Desk in Launceston and Hobart, and on the School's web site: http://www.cis.utas.edu.au/cisview/resources.jsp.

Students must take responsibility for the correct submission of their assignments. Students are expected to adhere to the following procedure for submission:

- Submitted files MUST be checked by the student to ensure that correct submission of the file has been undertaken.
- Students are expected to notify the Lecturer WITHIN TWO HOURS of submission if their files have not been submitted correctly.
- Students must take responsibility for safely backing up of their own files during the academic year to ensure that no files are permanently lost.

Extensions

Assessment items will not be accepted after the due date except under the conditions stated in the School policy on late assessment. http://www.cis.utas.edu.au/downloads/ExtensionPolicy.pdf (PDF - 100KB).

Review of Assessment and Appeals

1. It is expected that students will adhere to the following policy for review of any piece of continuous assessment.
   a. Within 5 days of the release of the assessment result, the student should request an appointment with the Lecturer. The student should be prepared to discuss specifically which section of the marking criteria they are disputing and why they consider the mark is inappropriate.
   b. Following this discussion, students may request a formal remark of the original submission (in accordance with Rule of Academic Assessment 111, clause 22.1). This remark will be undertaken, where practicable, by an alternative assessor.
2. Students may also request a review of the final result in a unit. The request and payment must be made within
Complaints Procedure

It is expected that students will adhere to the following policy for making any complaint or grievance directly related to a Unit:

a. In the first instance, students are to approach the Lecturer or Unit Coordinator concerned and arrange a time to speak with them about their concern.

b. If an issue remains unresolved, the student should approach the Head of School and arrange a time to speak with them about their concern.

If the School’s internal policy of complaints is unable to resolve an issue, students should consult Ordinance 8 Student Complaints for further direction, see http://acserv.admin.utas.edu.au/complaints_info.html

Formal Examination

The formal examination is conducted by the University Registrar. The 'Current Students' section on the university website contains information about the conduct of, and timetable for, formal examinations.

Final Grade

Passing grades will be awarded based on the AVCC guidelines:

- PP at least 50% of the overall mark but less than 60%
- CR at least 60% of the overall mark but less than 70%
- DN at least 70% of the overall mark but less than 80%
- HD at least 80% of the overall mark

In order to comply with the benchmarks set by the Faculty of Science, Engineering & Technology for distribution of grades in units, both the in-semester and examination marks that students obtain may be adjusted either upwards or downwards. See http://fcms.its.utas.edu.au/scieng/scieng/policies.asp for details of the Faculty Assessment Guidelines.