ISYS30221 Artificial Intelligence:
TOPIC A: Machine Learning Methods for Data Mining

I. Overview

The digital revolution has made data easy to digitally capture and inexpensive to store. Subsequently, many organisations and businesses are increasingly collecting data about various aspects of their operation, services, customers, suppliers and competitors with the hope of finding useful information with which to gain leverage or competitive advantage. As a result, **many organisations are storing between 20-100 terabytes of data *finding themselves data rich yet information poor*** with a recent study by IBM showing that 1:3 business leaders do not trust their data. This is likely to continue with IBM estimating that 2.3 trillion gigabytes of data is being created per day, expecting this to increase to 43 Trillion GB (40 Zetta bytes) by 2020. This has *led to an exponential demand (912% increase!) for professionals with expertise in managing and analysing big data sets*. This is particularly true for the following domains:

- **Business and Finance:** identifying and interpreting buyer/seller behaviour patterns, risk assessment data, product distribution patterns, foreign exchange rates, stock indexes and prices, interest rates, credit card data, fraud;
- **Government:** identifying behaviours and data patterns relating to: policy changes, tax evasion, benefit fraud, terrorism, social problems, law and order;
- **Internet:** extracting useful information to support personal or business decisions;
- **Health Care:** interpreting diverse diagnostic and treatment/patient care information stored by hospital management systems and biological/biometric data to inform patient care and drug development;
- **Telecommunication network:** Calling patterns and fault management systems.

The aim of this part of the module is therefore to ***introduce you to a number of 'machine learning' techniques to process and discover patterns in data***. Machine learning allows computational systems to adaptively improve their performance with experience accumulated from the observed data rather than to use pre-defined rules.

The content is themed around the notion of data as a resource; how it is acquired, prepared for analysis and finally how we can learn from it. A practical approach to teaching machine learning will be adopted and you will therefore use the freely available WEKA data mining package (www.cs.waikato.ac.nz/ml/weka) to put theory into practice. WEKA is full industrial-strength implementation of the techniques taught on this module and will enable you to better understand the techniques taught, appreciate their strengths and applicability to various problem domains.

The **essential reading** for this section of the module is as follows:

- Witten, I., Frank, E., Hall, M and Pal, C (2016), *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann; 4th edition (22nd Dec 2016), ISBN-13: 978-0128042915.

  NB we will make use of Witten's youtube videos (links given in this document and the lab sheets). You are strongly encouraged to watch ALL of the youtube videos recommended

You will also find the following textbooks very useful:

- Samarasinghe, S. (2007) Neural Networks for Applied Sciences and Engineering, First edition, Auerback Publishing, ISBN: 9780849333750

Online copies of both textbooks are available **for free** in the Resource List on NOW. Also, a number of supporting case studies and research papers will be made available on NOW for each of the taught topics. Note that professionals in the data mining field typically follow: http://www.kdnuggets.com/ - you too may find it useful too!

The content will be delivered over 9 weeks and organised into the following four parts:

Part 1. Introduction to Knowledge Discovery in Databases
Part 2. The Basic Components of Machine Learning Models
Part 3. Supervised Machine Learning Methods for Classification and Prediction
Part 4. Unsupervised Machine Learning Methods for Clustering

It is important to state that you will be exposed to a broad array of approaches to data analysis in a very short space of time, which will initially appear rather daunting! However, your assessment will enable you to focus your learning and thus develop a deeper understanding of a specific set of methods, models and techniques as appropriate to your chosen data set(s) and analytics task. You are therefore NOT expected to become an expert in every machine learning or data analytics technique!

The remainder of this document provides an indicative teaching and learning schedule for the 9-week teaching period for this section of the module. You should read it carefully and ensure you follow the directed reading. If you have any queries about this section of the module, then please contact me.

Jonathan Tepper

Dr Jon Tepper
Room ERD284 / Office hr: Tuesday, 1100-1230 / E: Jonathan.Tepper@ntu.ac.uk / T: 0115 8488363

II. Indicative Teaching and Learning Schedule

The indicative teaching and learning schedule for Topic A of the ISYS30221 Artificial Intelligence module is given in table 1 below. (Note that this schedule may change under exceptional circumstances. Every effort, however, will be made to notify you beforehand).

Uni Week	Week Comm.	Indicative Content and Reading		
		Lecture	**Laboratory**	**Core Reading/viewing**
11	09/10/17	**Part 1. Introduction to Data Analytics and CRISP-DM** a) The CRISP-DM process b) Getting to know your data	N/A	Witten et al (2016): pg 3-65 See NOW for additional research papers and case studies including: BBC Horizon programme: The Age of Big Data (watch it here…you will have to register and may have to search for the programme): http://bobnational.net/record/155036/media_id/158510) Also listen to the following Radio4 programme: http://www.bbc.co.uk/programmes/b03kqfzx ..and watch the Click programme on AI: http://www.bbc.co.uk/programmes/b06dst0x **Useful articles/sites:** https://www.youtube.com/watch?v=hZxnzfnt5v8 (types of data) http://changingminds.org/explanations/research/sampling/sampling_terminology.htm#frac (quick ref to relevant stats sampling terms) https://www.analyticsvidhya.com/blog/2017/09/6-probability-distributions-data-science/ (common distributions we should know)
12	16/10/17	**Part 2. The Basic Components of Machine Learning Models** a)Defining the learning problem b)Components of any machine learning system	N/A	Witten et al (2016): pg 43-88 (See NOW for additional research papers and case studies) http://www.youtube.com/watch?v=b4zr9Zx5WiE (Peter Norvig from Google, Machine learning hottest tech trends in 3-5 years) http://www.tutorialspoint.com/data_mining/index.htm (general introduction to data mining and machine learning) https://www.youtube.com/watch?v=yDLKJtOVx5c (Machine learning and its applications)
13	23/10/17	c) Model Fitting, Selection and Evaluation • Training and testing • Predicting performance • Cross-validation and bootstrapping • Confusion matrix / ROC curves • Calculating Recall and precision	1. The Explorer • Preparing and loading given data set(s) • Reviewing data filtering algorithms • Reviewing the Learning Algorithms in Weka – noting different types of algorithms (Bayes, Trees, Rules, Functions, Lazy, MI and MIsc)	Witten et al (2016): pg 161-202, 479-501 http://www.youtube.com/watch?v=hihuMBCuSlU (Model Selection with Cross Validation) https://www.youtube.com/watch?v=jiQamxz2ZcQ&nohtml5=False (Bias-variance trade-off) https://www.youtube.com/watch?v=Z5TtopYX1Gc (Sensitivity vs Specificity) http://www.youtube.com/watch?v=lHa1UYAxGxs (ROC curve #1) http://www.youtube.com/watch?v=sWAJsiVh1Gg (ROC curve #2) https://www.youtube.com/watch?v=vbpm9-WYuzU (Precision and Recall) https://www.youtube.com/watch?v=0zZYBALbZgg (Disproving through Hypothesis

Uni Week	Week Comm.	Indicative Content and Reading		
		Lecture	Laboratory	Core Reading / viewing
		• Evaluating numeric prediction (e.g. mean squared error (MSE), root mean squared error (RMSE), correlation coefficients)	• Pre-processing data and creating your own data sets in Arff format 2. The Knowledge Flow Interface • Getting started – add data, select attribute and tree method • Reviewing Knowledge Flow components: Visualisation and Evaluation • Configuring and connecting components • Running the experiment and interpreting the output	testing) https://www.youtube.com/watch?v=eyknGvncKLw (p values) https://www.youtube.com/watch?v=lwpobQmUTd8&nohtml5=False (confidence intervals) https://www.youtube.com/watch?v=fXOS4Q3nJQY (z scores) https://www.youtube.com/watch?v=0Pd3dc1GcHc&nohtml5=False (t-tests) https://www.youtube.com/watch?v=UmAJJtEo6cQ (z or t scores?) https://www.youtube.com/watch?v=1Ldl5Zfcm1Y&nohtml5=False (Chi squared test) https://www.youtube.com/watch?v=t2ryZyytW5w (two means t- test in Excel) https://www.youtube.com/watch?v=c68JLu1Nfkw&nohtml5=False (Evaluating numerical prediction models) https://www.youtube.com/watch?v=JxweGr03W9g&list=PLea0WJq13cnCZZ3sXVEZ2OE5CLeZUlCmm&nohtml5=False (Model Evaluation)
14	30/10/17	**Part 3. Supervised Machine Learning Methods for Classification and Prediction** a) Keep it Simple Stupid…probably! • Occams Razor • 1R - inferring 1 rule from data • Naive Bayes - Quick tour of probability theory - Bayes Theorem - Naive Independence assumption - Example classification - Limitations **Surgery Session #1 (Getting familiar with the coursework:**Mark sample coursework submissions**)**		Witten et al (2016): pg 96-105 http://www.youtube.com/watch?v=-8eSOmTPUbk (Intro to probability) https://www.youtube.com/watch?v=yi97YB8EyDg&nohtml5=False (Conditional prob) https://www.youtube.com/watch?v=3XL6w4_EKrg (Intro to Bayes Theorem) http://www.youtube.com/watch?v=wxmNvH7XBOA (Mario Kart and Bayes!)
15	06/11/17	b) Divide and Conquer with Decision Trees • What is a decision tree? • Learning from Claude Shannon: Father of the	3. The Experimenter • Getting started – add data and models to evaluate • Running an experiment • Analysing the results	Witten et al (2016): pg 105-112, 209-221 http://www.youtube.com/watch?v=wL9aogTuZw8 (overview of ID3 algorithm) http://www.youtube.com/watch?v=-dCtJjlEEgM (Lecture from Uni of Columbia) https://www.youtube.com/watch?v=z2Whj_nL-x8&noredirect=1 (The genius of Claude Shannon)

Uni Week	Week Comm.	Indicative Content and Reading		
		Lecture	Laboratory	Core Reading/viewing
		information age • Learning decision trees from examples - Quinlan's ID3 algorithm - Limitations and solutions - Applications	• Revisiting The Explorer • Training, testing and visualising different machine learning schemes	
16	13/11/17	c) Simple Linear Regression for Numerical Prediction • What is linear regression? • Simple linear regression - Ordinary least squares technique - Approach 1: using original data - Approach 2: using redefined data - Making predictions • Calculating correlation coefficients - Coefficient of determination (R2) - Pearson's coefficient of correlation (R) - Spearman's coefficient of rank correlation (Rs) • Evaluating the fit Other forms of linear regression **Surgery Session #2 (Share and discuss coursework ideas)**	4. Numerical Prediction • Reviewing numerical data sets • Using Weka for numerical prediction using simple linear regression • Evaluating what's been learnt	Witten et al (2016): pg 128-133 https://www.youtube.com/watch?v=ZkjP5RJLQF4 (Brandon Foltz Lec #1 - Basics) https://www.youtube.com/watch?v=iAgYLRy7e20 (Brandon Foltz Lec #2 - Algebra) https://www.youtube.com/watch?v=Qa2APhWjQPc&nohtml5=False (Brandon Foltz Lec #3 – Least Squares) https://www.youtube.com/watch?v=kHZBy1uVNnM&nohtml5=False (Brandon Foltz Lec #4 – Evaluating fit) https://www.youtube.com/watch?v=dQNpSa-bq4M&list=PLIeGtxpvyG-IqjoU8IiF0Yu1WtxNq_4z-&nohtml5=False (Brandon Foltz Lec on Multiple Regression) https://www.youtube.com/watch?v=Y2khrpVo6qI&nohtml5=False (Time series processing using ARIMA models)
17	20/11/17	d) Artificial Neural Networks (supervised learning) • Basic principles • The Perceptron • Problems of Linear Separability • Multi-layered Perceptrons (MLP) trained with Back-	5 & 5S Artificial Neural Networks • Basic matrix algebra for understanding artificial neural networks • Working with the perceptron for simple logic gates • Understanding the basics Back-propagation	Witten et al (2016): pg 260-272, 417-466 Samarasinghe (2007): 1-151; 195-220; 245-261. http://www.youtube.com/watch?v=DG5-UyRBQD4 (Understanding neural nets without the maths) https://www.youtube.com/watch?v=S3iQgcoQVbc&nohtml5=False (Perceptrons for simple classification) https://www.youtube.com/watch?v=0qVOUD76JOg&nohtml5=False (TED talk on neural nets)
18	27/11/17			

Uni Week	Week Comm.	Indicative Content and Reading		
		Lecture	Laboratory	Core Reading/viewing
		propagation - Applications **Surgery Session #3 (Getting familiar with coursework:** Remark sample coursework scripts**)**	- Getting back to Weka Using MLP with Back-propagation to classify and perform prediction (to be completed as homework if not during the lab).	https://www.youtube.com/watch?v=u5GAVdLQyIg and https://www.youtube.com/watch?v=IlmNhFxre0w (Engaging lectures on multi-layered perceptrons) https://www.youtube.com/watch?v=aVId8KMsdUU&list=PL29C61214F2146796&nohtml5=False (Advanced - derivation of the Backpropagation Learning algorithm for the interested student!)
19	04/12/17	**Part 4. Unsupervised Machine Learning Methods for Clustering** a) Simple K-means Independent study for the interested student: b) Kohonen's Self-organising Maps (SOMs)	6. Clustering in WEKA - Clustering a simple data set using K-means and other algorithms - Evaluating the clusters	K-means: Witten et al (2016): pg 141-156. https://www.youtube.com/watch?v=zHbxbb2ye3E&nohtml5=False (Fantastic illustration/visualisation of the K-means algorithm) For interested student (SOMs): Samarasinghe (2007): 337-436. https://www.youtube.com/watch?v=H9H6s-x-0YE&nohtml5=False (Understanding SOMs within 6 mins!) https://www.youtube.com/watch?v=HKj2ASG0DKQ&ebc=ANyPxKqz60EIWrF9Dg4V0Fzy-XE58j2vEEdXiIaGdMpyRpJ8fNFjbsQCkIYQbWzU_GqXyd3pX8tWDQGOT2Z7u9D7gfbPPX5rYQ&nohtml5=False (More detailed treatment of SOMs) https://www.youtube.com/watch?v=iWPhGKniTew (Lecture by the master himself, Teuvo Kohonen)

Table 1. Indicative content and reading schedule for Topic A. Machine Learning Methods for Data Mining. You should make every effort to view ALL of the recommended youtube videos.