Intelligent Systems in Business

David Barber

Intelligent Systems

- I have in mind application areas in which a more traditional 'build a solution' based on a full understanding of the problem is not feasible.
- These applications will typically have large databases.
- In essence, what we don't fully understand is made up for by having large amounts of data.
- The 'intelligence' is in extracting useful information from these by making models or summarisations of the data.

Course overview

Aims

- Provide an external viewpoint on the business applications of Intelligent Systems
- Stimulate a questioning and inquisitive approach to the field
- Material is topical and informative and presented to encourage discussion

Learning outcomes

- Ability to assess the effectiveness of solutions presented and to question them in an intelligent way
- Synthesise solutions to general open-ended problems covering material from the whole programme, tempered with information on commercial reality obtained from this course

Assessment

- Coursework (1 piece, 30%)
- Written Examination (2.5 hours, 70%)
- The examination rubric is: Answer THREE questions out of FIVE. All questions carry equal marks
- To pass this course, must obtain an average of at least 50% when the coursework and exam components of a course are weighted together

Coursework

- An essay of 5000 words (around 20 pages) on an application area of your choice.
- The essay topic must be approved by me before 22nd January.
- The deadline for hand in is 8 March at 12.00 – no electronic submission. The standard CS late submission rules apply.
- This is individual coursework. The standard CS plagiarism and collusion rules apply.
- You will also be asked to give a 20 minute presentation on the essay topic which will constitute 40% of the coursework mark.
- You will be assigned a date for the presentation. This will most likely be from the 11 March onwards, but depends on the invited speakers.

Essay contents

You must include in your essay:

- A detailed description of the application area, including motivations and background to the problem
- A discussion of the market for such an application, including companies that are active in this area.
- The history – what techniques have been tried in the past
- A summary of the strengths and weaknesses of available techniques
- You must include detailed technical descriptions of the techniques used.
- A discussion of the future outlook for this application area.

IS in the real world

- Finance (prediction)
- Automatic Speech Recognition
- Data mining
- Bioinformatics
- Healthcare
- Natural Language Processing
- Surveillance
- Games
- Dating
- Web
- Robotics
- Biometrics
- Military

Speakers

- **Tom Khabaza** (www.khabaza.com) *9 Laws of Data Mining*
- **Norman Casagrande** (last.fm) *How to organize digital music in a large social network*
- **Michael Mainelli** (Z/Yen Group) *Compliance architectures - The implications for machine learning*
- **Dimitrios Athanasakis** (National Institute for Medical Research)
- **Toby Mostyn** (meaningmine.com) *Extracting meaning from large amounts of unstructured data*
- **Alexander Korenberg** (Kilburn and Strode LLP) *Patents and machine learning*
- **Steve Poulson** (Scansafe) *Machine learning applied to (online) security*
- **Steven Barret** (Glaxo Smith Kline) *Drug discovery*
- **Janaina Mourao-Miranda** (Siemens) *Signal processing with EEG*
- **Thore Graepel** (Microsoft Research) *Online Gaming*
- Plus ShadowRobot.com and Infermed.com (to be confirmed)

Finance/Banking

- Stock price prediction
- Portfolio optimization
- Credit rating/ credit risk assessment.
- Fraud detection — credit card transactions
- Predicting bankruptcies

Speech Processing

- Automatic Speech Recognition
- Keyword detection (all voice transmissions monitored – 'Echelon'; Telegraph 9 Nov 2009)
- Voice coding (compression)

Music

- Signal Processing (compression)
- Music recommendation

Web

- Search : future – find me images like this; sound like this etc.
- The web as a knowledge source.
- Understanding the content of websites
- Product placement on websites (Touchclarity)

Natural Language Processing

- Text analysis
- Query understanding (search)
- Automatic parsing of addresses etc.
- Machine Translation
- Understanding relations between people/objects (Enron)

Visual

- Face Recognition
- Video surveillance
- Face detection (modern cameras)
- Scene understanding/3D reconstruction
- Human motion analysis (gait tracking/modelling)
- Satellite imagery analysis – automatic classification of land use

Gaming

- AI in games
- Understanding player ability (TrueSkill)

Marketing

- Customer credit, billing, and purchases were some of the first business transactions to be automated with computers, yielding huge amounts of data available for mining in search for knowledge that can improve marketing results or lower marketing costs.
- A typical example is direct mailing. A test mailing is made to a small subset of customer. A prediction is make as a function of the characteristics of the customer. This model can then be used to determine who should be included in the subsequent mass mailing and which offers should be included.
- Other examples can be found in customer relationship management (enhance the revenues of existing customers by tuning marketing messages) and preventing customer retention (identifying customers who are likely to switch to competitors).
- Causata.com – want to find causal relationships in data :‘ If we do x, the customer will cancel the contract’

Retail and Logistics

- Demand forecasting: In principle, these are standard time-series prediction problems. A nontrivial application has been developed for the prediction of single-copy newspaper sales (De Telegraaf - Just Enough Delivery, Smart Research).
- Predictions are needed on a daily basis for a huge set of individual outlets. The outlets can "learn from each other", e.g., by extracting typical demand features

Betting

- Betting : based on historical records can we beat the 'bookie odds'
- Analysis of betting (insider tips/fraud detection)
- Are there patterns in the history of betting that would suggest there is fraudulent activity

Military

- Tracking (radar – problems of multiple paths, bearing only tracking)
- Detection/analysis (is it a tank/person?)
- These analyses can be based on sound/vibration/visual cues

Data Mining

- Market analysis/Questionnaire
- Supermarket data analysis
- Where to place products in the store?
- What price to use?
- What is the optimal discount strategy – when should a good be reduced to 1/2 price?

Collaborative filtering

- Recommendation systems (Amazon.com, netflicks.com)
- Social network analysis

Healthcare

- Healthcare : understanding patient history
- Looking towards individually tailored treatments based on your genes and history/lifestyle (insurance!)
- Ageing society : monitoring of the elderly – changes in behaviour.

Crime/Security

- Understanding criminal networks
- Making networks based on disparate information sources

Robotics

- Industrial
- Help in an aging society (Japan)
- Military (Big Dog)

Demand Forecasting

- Newspaper sales
- suncream etc
- Electricity

Bioinformatics

- Understanding the genome
- Individual patient treatment
- Protein interactions
- Protein structure prediction ('folding')

Manufacturing

- The quality of a manufactured product often depends on the settings of many parameters. The exact relationship between these settings and the quality are often not well understood and too complex to describe with a physical or chemical model.
- Trained on examples yielding good and bad qualities, neural networks can provide a solution. Other applications of neural networks are in job shop scheduling and automatic inspection. In these control applications, neural networks are mainly used for function fitting to model (part of) the process one needs to control.

Health and medicine

- Detection of fraudulent insurance claims, risk assessment of clients, etc
- Automatic diagnosis of diseases.
- Expert systems in diagnosis (Promedas)
- Modelling of medical processes (the virtual human)

Energy and utility

- Prediction of energy demand is very relevant, both for large consumers who are often charged based on their peak energy usage, and for providers that have to anticipate upon extreme demands. Nonlinear time-series predictors.
- A quite different kind of application in this area involves the detection of likely sites for gas and oil deposits. Based on all kinds of measurements at test drilling sites, changes in the strata of rock, which relates to the presence of mineral deposits.

Challenges for Industry

- Machine translation
- More complete natural language understanding (search/knowledge acquisition)
- Speech recognition
- Visual scene understanding

Personal Consultancy Experience

- Company knows exactly what it wants but doesn't have the expertise in house
- Company knows what they want (but not necessarily have the same language to express it)
- Company doesn't know exactly what it wants (looking for an 'idea' to sell)

Academia and Industry

- Industrial interest can be driven by factors that they find appealing (not necessarily justifiable)
- For example : may be they are excited by 'genetic algorithms', 'neural networks' etc – it makes 'sense' to them since clearly nature is using these, so this must be a good thing
- Academia and industry are heavily influenced by current trends
- Academics often move more quickly from one toy-problem to another – more interested in methods than in outcomes
- Academics often are interested in 'optimal' solutions – industry in a robust solution that is cheap and effective

Real world data

- Real world data is often very messy
- Missing data
- Messed up data

Coding data

Categorical (unordered)

- May be transformed to a binary vector using 1-of-n encoding.
- Useful for algorithms such as k-nearest neighbours algorithm since the Euclidean distance between any two categories is the same.

Coding data

Ordinal

- As there is a natural ranking of the categories, it makes sense to code each category into a single real-valued number. E.g. we might have $cool = 0.1$, $cold = 0.5$, $warm = 0.8$ and $hot = 1.0$.
- The choice of the values between 0 and 1 might seem arbitrary, but if we have a number of examples drawn from a distribution, then we can work out the relative frequencies of each class and make a percentile coding. For example if we have 10, 40, 30 and 20 examples of cool, cold, warm and hot respectively, then we obtain the numbers given above.

Coding data

Ordinal

- A second method for dealing with ordinal data is to use *thermometer* coding. For n categories we take $n-1$ bits. For the first category we turn on 0 bits, for the second 1, and so on up to the nth category for which we turn on all $n-1$ bits.
- If we take a simple Euclidean distance between these vectors, we see that a pair of entries that are further apart on the scale will have a larger distance.
- Compare this to 1-of-n encoding, where the distance between any entries that are not identical is the same.

Coding data

Numeric

- For numeric data it would be possible to use this directly as an input. However, say that one of our variables is the height of a person. Should it be entered in metres or millimetres? For a k-NN classifier this will clearly make a huge difference unless we have scaling factors on each dimension to compensate. The usual choice is to rescale each variable x to a new variable $\tilde{x}$ having zero mean and unit standard deviation:

$$\tilde{x} = \frac{x - \overline{x}}{s}, \tag{1}$$

  where $\overline{x}$ is the mean of the variable (in the training data available) and s is the corresponding standard deviation.
- This ensures that all inputs have equal magnitude initially. It is also good practice in numerical algorithms to scale variables so that they are of $O(1)$ when possible.

Dimensionality reduction

- Some types of data can give rise to a large number of inputs. Examples include images (with up to millions of pixels), audio data (sampled at high frequency) and the output of spectrometers.
- Applying this data raw to a machine learning algorithm would lead to many problems (the parameters will be ill-determined. Also we do not believe that each entry in these kinds of data is independent; nearby pixels in images are strongly correlated.

Feature extraction

- In feature extraction we create some new features based on the original ones
- The creation of some hand-crafted features (e.g. the number of holes in a character image)
- Dimensionality-reduction mechanisms eg PCA.

Feature selection

- For feature selection, we could consider using all subsets of size k out of the original d features. There are $\binom{d}{k}$ combinations – too computationally expensive. Use forward or backward strategies instead.
- In forward selection, we start with no features. At each step, the next feature is added so that the joint predictive performance of all the features available so far is maximized.
- Alternatively in backwards selection, we start with all features and sequentially remove features to maximize predictive performance until a stopping criterion is satisfied.
- Stepwise strategies may not find the optimal combination of variables. If there are two variables which individually are poorly predictive, but together are very predictive, then a forward selection strategy may not select one of these features. Backwards selection suffers less from this problem but with a large number of variables it is undesirable as overfitting is likely to be a problem.

Reject option

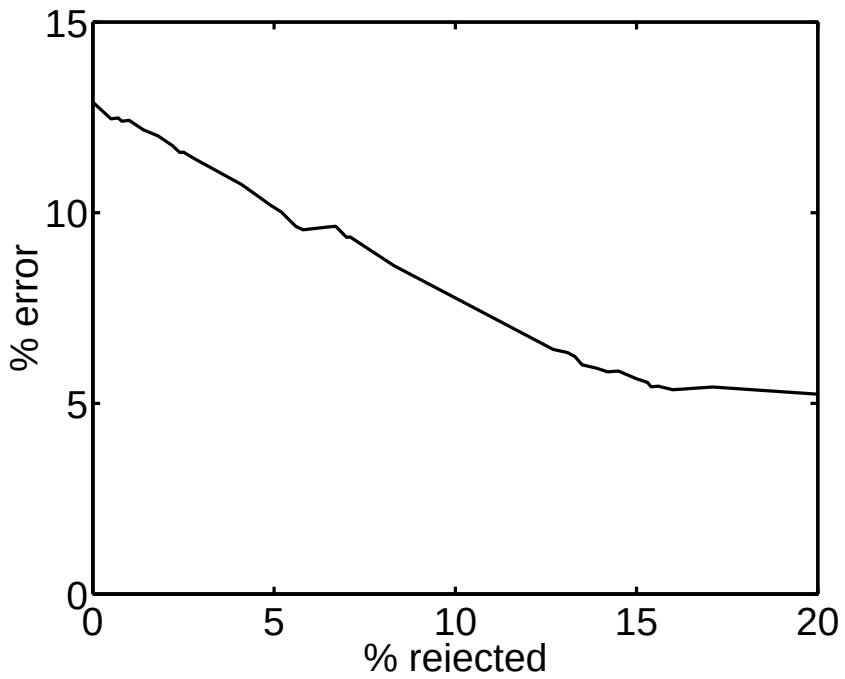
- If we have an estimate of $P(c|x)$, we may choose to reject some patterns, in the hope they are the difficult ones and that by throwing some out, the performance on the remaining patters is improved.
- For example the Post Office might require that an automatic postcode reader will deal with 95% of all letters, and that it obtains 99% or better correct classifications on those letters it does not reject.

Reject option

- If $\max_c P(c|x)$ is not very close to one, there is a chance that errors can occur. We can choose a threshold θ so that we carry out the classification iff

$$\max_c P(c|x) > \theta,$$

  otherwise we reject the pattern.
- As θ moves form 0 to 1, the fraction of rejected patterns increases. We can plot an error-reject curve:



Loss matrices

- The consequences of a misclassification can be more serious in some situations than in others.
- For example, if the machine does not flag a cancer when one is present (a false negative), this is much more serious than creating a false positive (which could then be eliminated in further testing).
- We need to predict the loss associated with the misclassification. Let L_{kj} denote the loss when a pattern from class $\mathcal{C}_k$ is assigned to class $\mathcal{C}_j$. In the cancer example, we might have

$$L = \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}$$

  where state 1 denotes "cancer" and state 2 denotes 'normal'.
- Given L, the optimal decision minimizes the expected loss.

Missing inputs

- If some inputs are missing, we have to ask if those attributes are *missing at random* (MAR), or if there is a systematic reason why they are not present.
- For example, a doctor would tend not to order particular tests if he felt that they would not be useful. In this latter case the deliberate non-measurement does convey information.
- For MAR data, let the input vector x be partitioned as $x = (x_p, x_m)$ where x_p denotes those attributes present, and x_m denotes the missing values. Then if we can model $P(x_m|x_p)$ we should average the predictions, weighted by this density.
- This is formally correct, but difficult to achieve in practice. A cruder version of this is to look for input patterns that have values on the present attributes similar to x_p, and to use then as the sample. An even cruder approach is to replace the missing values by their average values in the dataset.
- An alternative strategy is to create a "no value" label.

General advice

- Try to get as much training data as possible – this will improve both the performance of your model and your confidence in its predictions. If you get too much data so that training is very slow, figuring out which datapoints to discard is not too difficult.
- Black box methods should be treated with care. Usually, there are better solutions available if you set your mind to the issue of trying to make a reasonable model for the data.
- Try to avoid methods where you cannot easily figure out if the reason for poor performance is due to a fundamental poor choice of model, or in difficulties in the numerical application of the model. This is really the nightmare scenario since you don't know if you should just try another numerical approach (new optimisation procedure for example) or start fiddling with the parameters of your model.

General advice

- It is important to understand the assumptions behind any approach that you take. You are now the expert, and maybe your predictions are used in critical areas – tumour classification for example. If you don't understand well the fundamental model assumptions, there is really no science involved in your work, and you will not be able to easily defend or justify your position.
- Try to understand what it is that people want you to achieve in a machine learning problem. There are often different issues involved and the choice of method can be dependent on what the user ultimately wants.