Day 6: Machine Learning and Classification

Kenneth Benoit

Essex Summer School 2012

July 16, 2012

Continuum of Approaches so far..

- Purely qualitative approach - read the text and write up our judgement, (not very computer-aided)
- Thematic analysis - computer as bookkeeper
- Content analysis with coding
- Human-defined dictionary
- Automated dictionary
- Model similarity to known examples
- Choose a method that is (i) in keeping with your field and (ii) appropriate to your research question and data.

Machine Learning

- Relatively recent branch of a recent field (A.I.)
- Lots of published research and lots of practical applications
- Similar techniques to many social science models, but with a different terminology and philosophy
- Goal is to create algorithms which can make useful generalizations and predictions based on observed data.

Practical applications..

- Character recognition (postcodes, license plates)
- Medical and actuarial prediction and diagnosis
- Antenna design, circuit design, automated cars
- Product demand and market prediction
- Stock market and insurance modelling

For text analysis..

- Spam detection, language detection, translation, search expansion
- IBM Watson. Search engines, databases
- Sentiment analysis
- Speech recognition, natural language generation

Typical supervised learning framework

- Given a set of documents each belonging to a particular class
- Build a model based on the association between features of the documents and their class
- The model should be able to predict the class of new examples

Feature Value Matrix

- Generalization of term-document matrix
- Features might not be words, values might not be document frequencies
- All supervised machine learning algorithms define a similarity between new examples and previously seen examples for which the ‘correct answer’ is known.

Algorithmic approach

- Models viewed algorithmically (procedurally)
- Mostly depends on custom software
- Some public software packages exist: WEKA, Orange, NLTK, LibSVM

Classification vs Regression

- Regression in machine learning terms means trying to predict a value
- Classification means trying to predict a class
- Error for regression measured as a distance from the correct value
- Error in classification measured as proportion of examples classified correctly (accuracy)

An example - Naive Bayes Classification

- Choose the most probable class, given the data

Naive Bayes algorithm example

- Training Data:
- `The Dark Knight is really good`
- `I don't like the new Batman`
- `The Batman movie is good`
- `Bale is really bad in TDK`
- Test item:
- `I think the Batman film is good`

Nearest Neighbour algorithm

- Use values for features to map training examples to points in a space
- Map new example into the space and measure distance between the new example and each of the previous examples
- Give the new example the same label as its nearest neighbout, or take a vote among the labels of the K nearest neighbours.

Distance Measures

- Eucldian Distance measure
- Root of the sum of thesquared differences in each dimension (features)
- Cosine similarity - dot product divided by magnitude

Learning process

- Collect as much data as possible, as long as it is representative of the data that you want to apply the algorithm to.
- Divide the data into training, testing, and validation data
- Decide on features and text pre-processing
- Decide on methodology of implementation

Options for data collection

- Historical data
- Data from the same time in a different domain
- Manually generated data

Options for training and testing sets

- K-fold Cross validation
- Only seperate training and validation data (not testing)
- Divide training data into K portions
- Use one portion as testing data, others as training
- Alternate the portions, using each as testing data once
- Find average accuracy across all partitions
- If K = number of training examples, called "leave-one-out" cross validation

Measuring Error

- For regression and scaling, error can be measured qualitatively, or as a mean of the differences between predicted value and 'true' value
- For classification, error is measured as the proportion of correctly classified examples (accuracy)
- Accuracy can be misleading, depends on number of classes and distribution of examples among classes
- Baseline algorithms give meaning to accuracy figures
- Majority class - always predict the most frequent class
- Gibbs method - predict class with same probability of class distribution

Precision and Recall

- Same intuition as specificity and sensitivity earlier in course.
- Precision: $\frac{true positives}{true positives + false positives}$
- Recall: $\frac{true positives}{true positives + false negatives}$
- Accuracy: $\frac{Correctly classified}{Total number of examples}$
- F1: $\frac{Precision * Recall}{Precision + Recall}$

Amount of data required

- What is the cost of acquiring more data versus the benefif of reducing the error?
- Train on a small subset of your current available data and record performance on a test set
- Train on gradually increasing amount of data and graph relationship between size of the training set and accuracy
- Other costs - training time, equipment usage, testing time
- Use the most appropriate measure of error - are false positives and true negatives equally costly?

Feature selection and pre-processing

- Simplest approach for text - each word is a feature, its value for a given class is the sum of its frequency across each document in the class
- Other options:
- Aggregate frequencies across stems or lemmas
- Aggregate using a hand-compiled dictionary
- Aggregate known collocations or compound phrases
- Select features by learning correlation between feature choice and performance