Lecture 12: Outlook

Cognitive Systems II - Machine Learning

SS 2005

Outlook: Current Approaches and Applications

SVM, Active Learning, Data Mining

Machine Learning

We saw that ML is

- A sub-discipline of artificial intelligence, with large overlaps into statistics, pattern recognition, visualization, robotics, control, ...
- the study of computer algorithms capable of learning to improve their performance on a task on the basis of their own experience
- Often this is "learning from data.
- see talk of Gallagher "Current Trends in Machine Learning and Data Mining", Australian Virtual Observatory WS, 2003

Data Mining

- the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner
- Modern science is driven by data analysis like never before. We have an ability to collect and process data that is increasing exponentially!
- Increase in the size of datasets (in terms of observations and dimensionality)
- What to do when nr of dimensions $>$ nr of observations?
- Interest in data mining for non-numerical data.
- We are drowning in information and starving for knowledge (Rutherford D. Roger)

Steps in DM

- define problem
- collect data
- prepare data (use ML techniques)
- model data (use ML techniques)
- interpret/evaluate (use ML techniques)
- implement/deploy model

Machine Learning techniques are driven by the problems in data mining and provide some effective solutions.

Data Modeling and ML

Data modelling plays an important role at several stages in the scientific process:

- Observe and explore interesting phenomena. (unsupervised)
- Generate hypotheses.
- Formulate model to explain phenomena. (supervised)
- Test predictions made by the theory.
- Modify theory and repeat (at 2 or 3).

The explosion of data suggests that we need to (partially) automate numerous aspects of the scientific process.

Unsupervised Learning

- Given a set of d-dimensional data vectors $(X_1, \ldots, X_n)$, $X_i = (x_1, \ldots, x_d)$, build a model of the data to infer properties of the underlying distribution (process) that generated the data.
- Key problems:
  - Dimensionality reduction: developing algorithms that can reduce a dataset of hundreds or thousands of dimensions to just a few for visualization, while retaining as much of the "information" as possible in the original dataset.
  - Clustering of data - outlier detection: Identifying trends and/or anomalies in datasets.

Current Unsupervised Approaches

- Independent Component Analysis - decompose multivariate data with the aim of producing components that are as "statistically independent" as possible. Related to PCA and factor analysis.
- Gaussian mixture models for clustering - uses a semi-parametric probability density estimator that is trained iteratively on data. Implements a "soft" version of k-means.
- Self-organizing maps and topographic mapping - similar to clustering but where the cluster-centres are constrained to lie in a low-dimensional manifold (and so have a spatial relationship).

Supervised Learning

- Given a training set of pairs of input and output data vectors $\{(X_i, Y_i), \ldots, (X_n, Y_n)\}$, where the input values are thought to have some influence on the corresponding output values, build a model of the data that can predict the outputs of unseen (test) inputs.
- Key problems:
  Regression, classification, forecasting.
- Established approaches: Neural networks
  Decision-trees and rule-based classifiers
- Example applications in astronomy:
  Star/galaxy classification - on the basis of optical data. Photometric redshift evaluation. Noise identification and removal in gravitational waves detectors.

Current Supervised Approaches

- Support vector machines - uses insights from computation learning theory and geometry to produce predictors with powerful discrimination and good generalization.
- Ensemble methods - improving the accuracy of predictions by using multiple models and bootstrap sampling.
  - Example: boosting - incrementally constructs an ensemble of "weak" models, where each model is forced to concentrate on the mistakes made by previous models.

Current Supervised Approaches II

- Gaussian Processes - use the machinery of Bayesian inference to model data using stochastic processes.
  - All information is represented as a probability distribution.
  - Incorporates uncertainty associated with prior information and predictions made.
- Probabilistic graphical models - the model is a probability distribution where dependencies are explicitly encoded.
  - Generative models.

Current Supervised Approaches III

- Active learning:
  - Assume that data can be generated (measured or labelled) on demand - build a learning algorithm that learns on the basis of self-selected data points (queries).
  - Aims to reduce the amount of data required, training time of the model, or amount of data that must be manually labelled.

Current Supervised Approaches IV

- Semi-supervised learning:
  - Learning from both labelled (expensive, scarce) and unlabelled (abundant, cheap) data.
  - Aims are similar to active learning.
- Transductive learning:
  - All unlabelled data points belong to the test set.
  - The algorithm is able to take advantage of the spatial distribution of the data that the real-world generates (and that it will be tested on).

Other Trends

- Learning from Structured Data (Graphs)
- Learning Policies with RL
- Inductive Program Synthesis (Workshop AAIP at ICML 2005)