Data mining

Data to decisions

The idea

- Data mining/Machine learning is the process of detecting and exploiting patterns in "large" quantities of data
- Much of the big data talk deals with overcoming physical capacities:
  1. Storage
  2. Data preparation
  3. Bringing the needed computing power to bear
  4. Locating the appropriate data-management software
  5. Efficient (feasible) algorithms…
- We will focus on the application side

Data mining vs. statistical inference

- In traditional statistics (really, in science) we formulate a question and then collect the appropriate data or design the appropriate experiment
- Data mining, in fact, is viewed as a bad thing in science
- Data mining starts with data and gropes around for exploitable patterns
- It can be given solid foundations too, those come from the ever more active field of *machine learning*

Examples of data mining outcomes

1. Classification: learning how various attributes maps to a defined output
2. Association: discover relationships/connections between various items in large data (items that tend to be purchased together, e.g.)
3. Clustering (=unsupervised classification): group 'similar' data objects together without reference to a specific target variable
4. …

Classification

- Consider a collection of *records* (x, y) where x lists observable attributes (age, income, …) and y is a target variable of interest (buy or not, e.g.)
- Classification applies mostly to binary targets
- We are trying to learn how x impacts y
- This is something we can exploit for forecasting purposes, promotion strategies, churn prevention…

There are many ways to skin this cat

1. Estimate a categorical regression
2. Non-parametric regressions
3. Boosting methods
4. Neural networks
5. Decision trees
6. …

Decision trees

- Each attribute contains some information on the target variable
- A *strong attribute* is one where for a given value of that attribute, the target value tends to be uniform
- Why not start with the most informative attribute first, then go for the attribute that adds more information and so on and so forth…
- But if we keep going until we run out of attributes, we will obviously overfit since any attribute contains some information in sample
- Solution: *prune* the tree once finished, or cross-validate
- A decision tree method is *a splitting algorithm* together with a *stopping rule* or a *pruning algorithm*

Strong attributes and entropy

- Since the target is categorical, any subset of observations is a probability distribution over the possible values of the target
- If, say, we are looking at a binary $\{0,1\}$ target, the least information we have is a 50-50 split over those two choices
- How much information we have before splitting, more generally, is the distribution's entropy (see chapter 2 for a refresher)
- A strong attribute is an attribute where entropy given the value of that attribute is much smaller than prior to the split
- In plain English, in subcategory of that attribute tend to be associated with similar target values

But enough chit-chat, it's time to looks at some examples