Data Mining and Statistics: Introduction

Prof. Rafal Kustra

Public Health Sciences

September 19, 2016

Data Mining - my take

Definition

A set of activities to learn generalizable insights from large dataset not collected under strict statistical design principles

Definition 2

2- Statistical analysis of heteregenous dataset focused on prediction with loosely specified hypothotesis to test

Data Mining Features

What's different?

- Focus is on predictive power - but intepretability is still important (no black box)
- Many potential biases in dataset - not usually collected under controlled conditions
- A lot more of categorical measurements - including outputs of interest (classification)
- Data usually large in some dimensions
- Relational data sets
- Data preparation - paramount

Some DM problems

Public Health and Web Queries

- Google Flue Trend: correlate past google search terms with flu doctor visit reports from CDC
- Use IP geolocation to regionalize (state-wide and better)
- Use as a near real-time flu monitoring tool

Gene Function Prediction

- Biological Processes (about 10k) categorized in a graphical structure (Gene Ontology)
- Few hundred microarray experiments with 6 thousand yeast genes
- for about 4k genes some functionality known
- Use both the uArray dataset and Gene Ontology structure to predict functions of unknown genes

Models and data

- Data collected represents a simplified picture of the reality
- In data mining world one approaches a previously collected data to formulate and/or answer some questions
- Statistical models are used to both simplify the data *and* elicit answers (or crystalize questions)
- Most statisticians would say that no model is *right*, but some models are better than others and some models can be spectacularily wrong

George Box

All models are wrong but some are useful

Supervised and Unsupervised Models

Supervised

Has a "guiding" response, y, and a vector of predictors, x.
For continuous Y, seeks some kind of regression function:

$$\mathrm{E}(Y|X = x)$$

or class-density functions, for discrete Y:

$$\mathrm{P}(Y = y_k|X)$$

Unsupervised

Only "predictors", no response. Want to find "structure" in predictors
Usually connected to density estimation in some way, although may only be interested in some aspects of the density. Clustering, PCA, hot-spot analysis are examples.

Some supervised statistical models

- Linear regression:

$$Y_i = \alpha + \beta_1 x_{i1} \ldots + \beta_p x_{ip} + \epsilon_i$$

- Logistic regression:

$$P(Y_i = 1 | X_i = x) = \frac{1}{1 + e^{\alpha + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}$$

- Classically we are interested in hypothesis test on β_k for some X_k of interest:
  - Does distance from major hopsital (X_k) affect the length of delay to surgery (Y)?
  - Does treatment by new drug (X_k: binary) improve the odds of survival ($Y = 1$) compared to old drug?
- But both of these can - and often are - be used in prediction context

Some unsupervised research hypothesis

- What are different clusters of patients waiting in ER room of specific hospital?
- Are there few patterns of similar expressions among 6000 genes studies in 60 leukemia subjects?

Big Data: 4Vs

- Big Data is a newer take on Data Mining: with increased emphasis on Internet
- People talk about 4Vs of Big Data:

Volume This is obvious: lots of data (terrabyte becomes a bit small of a unit)

Velocity The data comes in constantly: from social networks, search engines, apps (VuZe now part of Google Directions), wearable sensors; but also from classical sources which have been internetifies, say medical tests and precscriptions etc

Variety The data is not one nice matrix. It can be phrases, images, digitized medical records. It can be a time stream or single measurment per subject. It can be aggregated (city, census track) or individual based.

Veracity Data is noisy AND noisily collected.

Big Data and Biostatistics

- Some things do not change: understand the data, understand potential biases, understand the question we want to answer with the data and limitations
- Some approaches and models will remain useful but may need extending
  - Linear regression does not have to be linear (same for logistic or Cox model or many other classical models)
  - Proper display of data can be worth a thousand models but with 4V data we may need to work harder to capture the underlying model
- But some data mining approaches may be very useful
  - Predictive analysis and validation
  - Feature construction
  - More complex learning that capture nonlinearities and more potential interdependencies (interactions)