Why do statisticians "hate" us?

David Hand, Heikki Mannila, Padhraic Smyth

> "Data mining is 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."

Results of data mining are models or patterns.

observational vs. experimental data

> data mining uses data that has usually been collected for some other purpose. Ex. retail data (record of purchases in a store)
>
> The data mining process may not be connected to the data acquisition process.
>
> In statistics data is often collected to answer specific questions
>
> Data mining sometimes referred to as secondary data analysis

Large data sets: If data sets were small we would use statistical exploratory analysis

Large to a statistician - few hundred to a thousand data points
Large to a data miner - millions or billions of data points

Exceedingly large datasets have problems associated with them beyond what is traditionally considered by statisticians

Many statistical methods require some type of exhaustive search. As datasets become larger wrt to records and variables, search becomes computationally more prohibitive (looking at all subsets of variables requires search of 2^p - 1 sets)

Increase in number of variables leads to *the curse of dimensionality*

curse of dimensionality - exponential rate of growth in the number of unit cells in the data space as the number of variables increases.

Ex. 1 binary variable - can take on a value of 0 or 1, hence two "cells" to get a good estimate of the variable let's say we need 10 observations per cell. 20 observations in all. 2^1 X 10.

two binary variables you need 40 observations. or 2^2 X 10.
10 binary variables you need 10240 observations or 2^{10} X 10.

20 variables 10485760 observations….

What if we have variables that take on multiple values? Things get MUCH worse.

Large data sets of issues with the access.
Statistical viewpoint:

John Elder, Daryl Pregibon, *A Statistical Perspective on Knowledge Discovery in Databases*

<u>A brief history if statistics relative to KDD</u>

1960's

"The robustness era freed statisticians of the shackles of narrow models depending on unrealistic assumptions (e.g. normality)"

Early 1970's

- Exploratory Data Analysis (EDA) - statistical insights and modeling are data driven
- Arguments against: Don't bias the hypothesis and the model
- Arguments for: By looking at the data and summaries of that data we can choose richer statistical models
- Statisticians could "look" at the data BEFORE they created a model!
    - statistical models could break the data into structure and noise
    - *data = fit + residual*
    - look at the residual and see if there was any more "fit" that can be done
    - iterative process
- Visualizations - graphs, pictures are worth a thousand words!! Humans can visualize relationships better than any program (still true!!)

- Data descriptive methods were more acceptable as opposed to mathematically based methods
    - data reexpression - log(age) vs. using raw age value
    - ignore outliers so to get a better description of the major portion of the data

Late 1970's

- normal theory linear model extended to generalized Linear Models which included probability models (Nelder, Wedderburn, 1974: McCullagh, Nelder, 1989)
- EM algorithm (Expectation Maximization) - ways to solve estimation problems with incomplete data. Complete data may also benefit from being treated as a missing data problem.
- Bishop Fienberg, and Holland - analysis of nominal or discrete data using loglinear models.

Early 1980's

- Elimination of low order bias of an estimator by using resampling (jack knife)
- technique was generalized to sampling without replacement from data (bootstrap)
- It allowed for the determination of the error in the estimate
- focus on estimating precision of estimators rather than bias removal

Late 1980's

- Use of scatterplot smoothers or sliding windows
- yielded models with global nonlinear fits
- applications in classification, regression, discriminaton

Early 1990's

- focus shifted from model estimation to model selection
- Many candidate models can be considered for data
- hybrid models, formed from multiple "good" models
    - yields models that are more accurate
    - reduced variance

Data Mining, Data Dredging, Snooping, Fishing

If you searched long enough you would always find some model to fit a data set arbitrarily well

- Two contributing factors:
    - complexity of the model
    - size of possible models
- If models are flexible wrt size of available data, then we can fit data arbitrarily well
- But we need to be able to generalize so that we can model new data - Avoid overfit
- Therefore we need simpler models
    - Occam's Razor - given a choice of solutions, the simplest may be the best

Extreme case - Model = Data

- Perfect Fit - Yes!
- Generalize Well - No!
- Interesting - No!

Given a simple model, consider a billion of them. One may fit!!

Ex. Predict a response variable *X* from a predictor *Y*

Given a very large set of variable, $X_1, X_2, X_3, \ldots, X_p$ where *p* is large, then there is a high probability of finding an *X* that correlates with *Y* even if no real correlation exists in the domain.

Ex. When a team from a particular football league wins the super bowl, then a leading stock market index goes up.

Leinweber: perfect prediction of S & P 500 as a function of butter production, cheese production and sheep populations in Bangladesh and the US.

No easy solutions. One strategy - split data set so that model is built on part of data and tested on another (cross validation)

Best soln - See if what you produce makes sense!! - To who? To the domain expert/user.