Visualizing and Exploring Data

- The Nature of Data Sets
- Summarizing Data
- Displaying single variables
- Displaying two or more variables
- Projection methods

Reading: HMS, chapter 3, Supplement to LfD Visualization notes

The Nature of Data Sets

- n objects (or cases, records etc)
- each with p attributes (features, fields, variables)
- A n x p data matrix
- Cf structured data (e.g. text, graphs)
- Attributes can be
  - Nominal (Categorical, Ordinal)
  - Numeric

Summarizing Data

- Measures of location: mean, median (for each attribute, if numerical)
- Measures of dispersion (variability)
  - variance
  - range (max-min)
  - inter-quartile range
- Covariance, correlation matrix
  \[ r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}}\sqrt{c_{jj}}} \]
Visualization

Objectives
- Present data in graphical form
- Allow user to actively “explore” the data (cf Exploratory Data Analysis, Tukey 1977)
- To process the data to aid exploration

Displaying single variables
- Histogram. Use “too many bins”: maximize information seen.
- Kernel smoother with width $h$

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

with $\int K(x)dx = 1$. Gaussian kernel is a common choice
- Dot plot (xgobi)
- Box-and-whiskers plot

Look for outliers, multimodality etc. See plots in Figs 3.2, 3.3 and 3.5 from HMS. Note suspicious 0s in Fig 3.2.

Displaying two or more variables
- Scatterplots (weka, xgobi) E.g. HMS Fig 3.13
  - Beware of overprinting …
  - Brushing
  - Spin plots
- Icons. E.g. HMS Fig 3.15
- Parallel coordinates

Projection methods
- Project multivariate data into 2 or 3 dimensions
- E.g., PCA projection: transform data into space defined by principal components, keep first 2 or 3.
- Classical scaling: given distance between data points in original space, find points in low-dimensional space where distances are similar.
Classical Scaling

Figure: Classical scaling solution to representing 28 world cities on a two-dimensional map, given only their intercity distances.

Projection methods

- Project multivariate data into 2 or 3 dimensions
- e.g., PCA projection: transform data into space defined by principal components, keep first 2 or 3.
- Classical scaling: given distance between data points in original space, find points in low-dimensional space where distances are similar.
- turns out these are related! (see LfD notes)
- If we have class labels, we can do more than just PCA—Canonical Variates considers variance within as well as across classes.
- PCA looks for a projection that maximizes variance. We can look for projections that maximize other measures of interestingness (e.g. non-Gaussian) \( \Rightarrow \) projection pursuit

Projection pursuit

Diaconis and Freedman (1984) proved that for most high-dimensional clouds, most low-dimensional projections are approximately normal. Implies (?) : non-normal projections are interesting (multi-modal, skew).

How to find interesting directions: \( y = \mathbf{a}^T \mathbf{x} \), with constraint \( \mathbf{a}^T \mathbf{a} = 1 \).

Estimate \( f_\mathbf{a}(y) \) using Parzen windows, and calculate an index \( Q(f) \) which quantifies non-normality, e.g. (Huber, 1985)

\[
\frac{1}{2} \log(2\pi e\sigma^2(y))
\]

or kurtosis

\[
E[y^4] - 3(E[y^2])^2
\]

Gradient based search to optimize \( Q \)

Projection pursuit is available with xgobi

Visualization: Heuristics

- Go for high information 2D visualizations. 3D visualizations should only rarely be used if there is no other way.
- Proactively (on the basis of previous visualizations) select data subsets to visualize.
- You must provide a potential explanation for any anomaly: you must never let anomalies pass you by. Dig deeper.
- Use your visualizations to inform potential models. Use your potential model to direct your visualizations.
- Expect problems in your data. Go in search for them. Know your data inside out before moving on.
- This is the cheapest and most informative stage of data mining.
- Failure to properly know your data will come back and bite you later on.
- Uses the power of eye/brain to find structure in data
- Opposite end of spectrum from formal model building
- Help to find unexpected relationships and to identify outlier etc
- Data-driven hypothesis generation