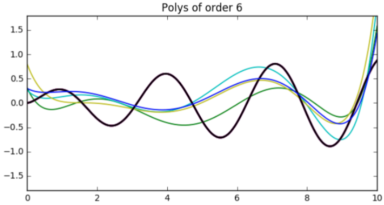
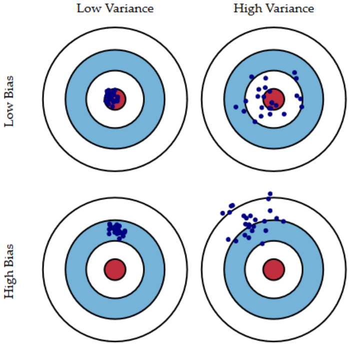
FIT1043 Introduction to Data Science

Module 5: Data Analysis Process

Lecture 10 - Part I

Discussion: Bias Variance





- Bias measures how much the prediction (averaged over all data sets) differs from the desired regression function.
- Variance measures how much the predictions for individual data sets vary around their average.

Unit Schedule: This Week

Module	Week	Content
1.	1	Overview and look at projects
	2	(Job) roles, and the impact
2.	3	Data business models / application areas
3.	4	Characterising data and "big" data
	5	Data sources and case studies
4.	6	Resources and standards
	7	Resources case studies
5.	8	**Data analysis theory**
	9	**Regression and decision trees**
	10	**Data analysis process**
6.	11	Issues in data management
	12	Guest Speaker and Exam Info.

Preprocessing Data
For building a Predictive Model

normalising features
imputing missing values

Normalisation

- Preprocessing step in building a predictive model
- Scale to fall within a small, specified range, e.g., [0,1]

Imputation

ID	Age	Amount	Duration	Job	Housing	Marital	Default
001	43	$200,000	240	A	apartment	yes	no
002	27	$150,000	280	A	apartment	no	?
003	?	$180,000	240	B	house	yes	no
004	42	$200,000	240	?	apartment	yes	no
005	31	$300,000	240	C	house	yes	no

- here we have the housing loan prediction problem
- record 002 has the target variable (*Default*) missing
  - cannot be used by standard learning algorithms
- record 003 has *Age* missing, record 004 has *Job* missing
  - if we "fill in" the missing variables using imputation then these records can be used

Theory of Data Analysis
Characterizing Learning

broad characterisations for general discussion

Characterizing Learning

Prediction: Is the task a simple prediction?

Dynamic: Does the task repeat over space or time? (GPS, game playing)

Missing data: Do some of the variables missing have missing data? (note they cannot be 100% missing)

Latent variables: Are there latent variables? *e.g.*, a segmentation task. Note the target variable for a prediction task cannot be latent.

Optimisation: Does evaluation/prediction require optimisation *after* statistical inference (*i.e.* after prediction)?

latent variable ::= variable whose value never appears in any data

Data Analysis
What is Hard?

The Hardest Parts

See blog *"The hardest parts of data science"* by Yanir Seroussi 23rd Nov. 2015.

Model fitting: core statistics/machine learning – not usually hard (*e.g.,* many use R as a black box for this)

Data collection: can be critical sometimes, but often more routine

Data cleaning: can be a lot of work, but often more routine

Problem definition: getting into the application and understanding the real problem can be hard

Evaluation: what is measured? should multiple evaluations be done? can be hard

Ambiguity and uncertainty: invariably these occur and we need to live with them; can be hard

Tools for the Data Analysis Process (ePub section 5.4)

popular software and prototyping

Common Software

access: SQL, Hadoop, MS SQL Server, PIG, Spark

wrangling: common scripting languages (Python, Perl)

visualisation: Tableau, Matlab, Javascript+D3.js

statistical analysis: Weka, SAS, R

multi-purpose: Python, R, SAS, KNIME, RapidMiner

cloud-based: Azure ML (Microsoft), AWS ML (Amazon)

KDnuggets on the R vs. Python debate

Scripting Languages

see Wikipedia entry *scripting languages*:

- no formal or universally agreed definition
- often interpreted and are high-level programming languages
- automating tasks originally done one-by-one by hand
- also, extension language, control language

e.g. bash, Perl, Python, R, Matlab, ...

Data Analysis (ePub section 5.7)

general considerations about data analysis

Data Analysis Case Studies

Google flu trends

Google Flu Trends

Google Flu Trends (2min YouTube video)
Google Flu Trends (PBS) (2min YouTube video)

- U.S. Centers for Disease Control and Prevention (CDC) and the European Influenza Surveillance Scheme (EISS) provide data with 2 week lag
- CDC has 9 surveillance regions in US and report "influenza-like illness" (ILI) visits weekly
- Google researchers
  - selected top 45 queries that predicted *ILI visits* across regions in 2003-2008
  - built a linear model on these 45 queries to predict *ILI visits* 2003-2007
  - tested it against 2007-2008 data

Google Flu Trends, cont.

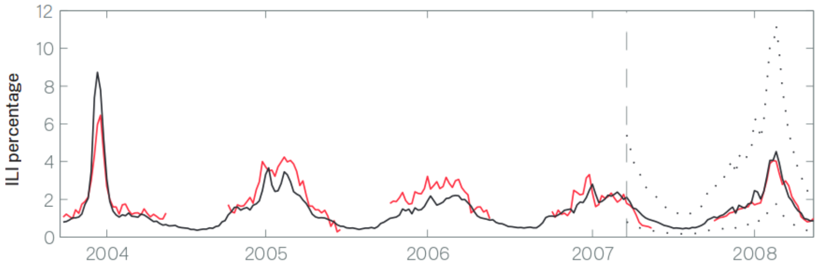


Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.

Google Flu Trends: Critique

see *Science* March 2014,
"The Parable of Google Flu: Traps in Big Data Analysis"

- The stability of logs is unclear (Google's search engine is evolving)
- lack of reproducibility (Queries used were not disclosed)
- Google could have augmented their query log signals with CDC's historical count data and made their predictions more robust.

More Case Studies

- See lecture slides Week 10 - Part II for more data analysis case studies.