Data Mining and Exploration

Michael Gutmann

michael.gutmann@ed.ac.uk
http://homepages.inf.ed.ac.uk/mgutmann

Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh

19th January 2017
Data

Oxford dictionary:
- Plural of datum
  - From Latin: dare, to give; datum: something given
  - A piece of information
- Facts [...] collected together for reference or analysis
- Things [...] making the basis of reasoning or calculation
“Data! Data! Data!” he cried impatiently.

“I can’t make bricks without clay”
Data sources

- Scientific measurements
- Business records
- Medical tests
- Paying by credit card
- Using the mobile phone
- Social media
- Machines
- ...
Scientific data

Large Hadron Collider:

- Particles collide at high energies, creating new particles that decay in complex ways
- The raw data per collision event is around one MB.
- About 600 million events per second.
  ⇒ 600 terabyte of data per second

Source: https://home.cern/about/computing
Human generated data

On a single day

- 500 million tweets
- 4.3 billion Facebook messages
- 6 billion Google searches
- 205 billion emails
- . . .

Source: https://www.gwava.com/blog/internet-data-created-daily
Machine generated data

- Airplane engine: 5,000 sensors, 10 GB of data per second
- Internet of Things

Data mining ≈ data analysis ≈ data science

First sentences from corresponding wikipedia pages:

- The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use

- Analysis of data is a process of [...] with the goal of discovering useful information, suggesting conclusions, and supporting decision-making

- Data science [...] is an interdisciplinary field about scientific processes and systems to extract knowledge or insights from data in various forms [...]

Michael Gutmann
DME
Data mining $\approx$ data analysis $\approx$ data science

In short:

- Data $\rightarrow$ knowledge
- Evidence $\rightarrow$ conclusions
- Pieces of information $\rightarrow$ actionable information
- The process of “making the bricks out of the clay”
Data analysis as statistical inference

Given a data generating process, what are the properties of the outcomes (the data)?

Based on Figure 1 of *All of statistics* by Larry Wasserman
Data analysis as statistical inference

Given a data generating process, what are the properties of the outcomes (the data)?

Given the outcomes (the data), what can we say about the process that generated them?

Data are a realisation of a random vector $\mathbf{x}$ with some probability distribution that we don’t know.
Data analysis process

Get (raw) data
- understand the sampling process
- any biases?
- feedback loops between data analysis and collection?

Sanity checks
- where are you now?
- what do you want to do?
- constraints?

Objectives and key results

Exploratory data analysis
- become familiar with the data
- spot unexpected properties
- anomalies, outliers, missing data?

Prep data for further analysis
- merge data sets, reformat
- select/exclude data
- provide clear rationale for selection/exclusion

Build and fit model
- generalisation is the goal
- choice of evaluation metric
- choice of hyperparameters

Summarise, visualise results
- can you tell a simple and coherent story?
- what makes sense, what not?
- limitations, uncertainties?

Deploy the product / Communicate findings

- merge data sets, reformat
- select/exclude data
- provide clear rationale for selection/exclusion

- generalisation is the goal
- choice of evaluation metric
- choice of hyperparameters

- can you tell a simple and coherent story?
- what makes sense, what not?
- limitations, uncertainties?
Plan for DME

Lecture 5
- understand the sampling process
- any biases?
- feedback loops between data analysis and collection?

Get (raw) data

Sanity checks
- become familiar with the data
- spot unexpected properties
- anomalies, outliers, missing data?

Objectives and key results
- where are you now?
- what do you want to do?
- constraints?

Lectures 1-3

Exploratory data analysis
- merge data sets, reformat
- select/exclude data
- provide clear rationale for selection/exclusion

Prep data for further analysis
- generalisation is the goal
- choice of evaluation metric
- choice of hyperparameters

Build and fit model
- can you tell a simple and coherent story?
- what makes sense, what not?
- limitations, uncertainties?

Summarise, visualise results

Deploy the product /
Communicate findings

Presentations
Mini-project