Lecture 1: Introduction

Felix Held, Mathematical Sciences

MSA220/MVE440 Statistical Learning for Big Data

25th March 2019

CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

What is Big Data?

What is Big Data?

- Is it just a buzz word?
- Is it a cure to everything? See e.g. [1]
- Big Data - Big Problems?
  - Big Data does not mean correct answers, see e.g. [2]
  - Privacy concerns, see e.g. [3]
  - Big Data is often not collected systematically, see e.g. [4]
- It's a huge topic in science! Almost 5 million hits on Google Scholar.

[1] `https://www.businessinsider.com/big-data-and-cancer-2015-9?r=US&IR=T&IR=T`

[2] Lazer et al. (2014) The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343 (6176):1203–1205. DOI 10.1126/science.1248506

[3] `https://www.nytimes.com/2018/03/22/opinion/democracy-survive-data.html`

[4] `https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0#axzz2yQ2QQfQX`

So Big Data is about size?

Yes and no.

Note that *size* is a flexible term. Here mostly:

- Size as in: *Number of observations*

  Big-n setting

- Size as in: *Number of variables*

  Big-p setting

- Size as in: *Number of observations **and** variables*

  Big-n / Big-p setting

Is this all?

The Four Vs of Big Data

Four attributes commonly assigned to Big Data.

Volume Large scale of the data. Challenges are storage, computation, finding the interesting parts, ...

Variety Different data types, data sources, many variables, ...

Veracity Uncertainty of data due to convenience sampling, missing values, varying data quality, insufficient data cleaning/preparation, ...

Velocity Data arriving at high speeds and need to be dealt with immediately (e.g. production plant, self-driving cars)

See also `https://www.ibmbigdatahub.com/infographic/four-vs-big-data`

How does statistics come into play?

Statistics as a science has always been concerned with...

- sampling designs
- modelling of data and underlying assumptions
- inference of parameters
- uncertainty quantification in estimated parameters/predictions

Focus is on the last three in this course.

Statistical challenges in Big Data

- Increase in sample size often leads to increase in complexity and variety of data (p grows with n)
- More data $\neq$ less uncertainty
- A lot of classical theory is for fixed p and growing n
- Exploration and visualisation of Big Data can already require statistics
- **Probability of extreme values:** Unlikely results become much more likely with an increase in n
- **Curse of dimensionality:** Lot's of space between data points in high-dimensional space

Course Overview & Expectations

A clarification upfront

This course focusses on statistics, not on the logistics of data processing.

- Understanding of **algorithms**, **modelling assumptions** and **reasonable interpretations** are our main goals.
- We will focus on well-understood methods supported by theory and their modifications for big data sets
- No neural networks or deep learning. There are specialised courses for this (e.g. FFR135/FIM720 or TDA231/DIT380).

Themes

- Statistical learning/prediction: Regression and classification
- Unsupervised classification, i.e. clustering
- Variable selection, both explicit and implicit
- Data representations/Dimension reduction
- Large sample methods

Who's involved

Felix Held, felix.held@chalmers.se

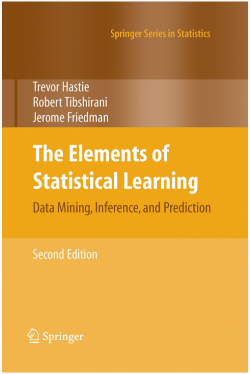
Rebecka Jörnsten
jornsten@chalmers.se

Juan Inda Diaz
inda@chalmers.se

A course in three parts

1. Lectures
2. Projects
3. Take-home exam

Course literature



Hastie, T, Tibshirani, R, and Friedman, J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer Science+Business Media, LLC

- Covers a lot of statistical methods
- Freely available online
- Balanced presentation of theory and application
- Not always very detailed. Other suggestions on course website.

Projects

- **Five (small) projects** throughout the course
- **Purpose:**
  - Hands-on experience in data-analysis
  - Further exploration of course topics
  - Practice how to present statistical results
- You will work in groups and have at least one week per project
- Projects will be presented in class
- Attendance (and presenting) of project presentations is mandatory to be allowed to take the exam
- More information next week

Exam

- **Take-home exam**
- **Structure:**
  - 50% of the exam/grade: Revise your projects individually
  - 50% of the exam/grade: Additional data analysis/statistical tasks
- Exam will be handed out on 24th May
- Hard deadline on 14th June

Statistical Learning

Basics about random variables

- We will consider **discrete** and **continuous** random quantities
- **Probability mass function (pmf)** $p(k)$ for a discrete variable
- Probability density function (pdf) $p(\mathbf{x})$ for a continuous variables

Two important rules (and a consequence)

Marginalisation

For a joint density $p(x, y)$ it holds that

$$p(x) = \sum_y p(x, y) \quad \text{or} \quad p(x) = \int p(x, y)\,\mathrm{d}y$$

Conditioning

For a joint density $p(x, y)$ it holds that

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

Both rules together imply **Bayes' law**

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Expectation and variance

Expectations and variance depend on an underlying pdf/pmf.

Notation:

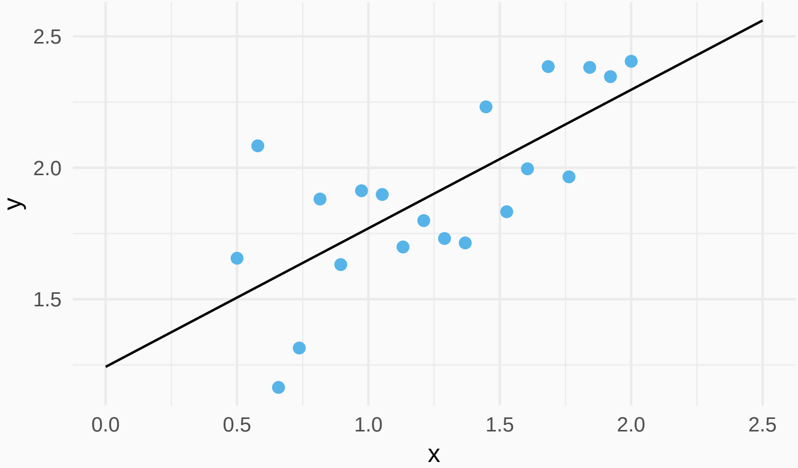
- $\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)\,\mathrm{d}x$
- $\mathrm{Var}_{p(x)}[f(x)] = \mathbb{E}_{p(x)}\left[\left(f(x) - \mathbb{E}_{p(x)}[f(x)]\right)^2\right]$

What is Statistical Learning?

Learn **a model** from **data** by minimizing **expected prediction error** determined by a loss function.

- **Model:** Find a model that is suitable for the data
- **Data:** Data with known outcomes is needed
- **Expected prediction error:** Focus on quality of prediction (predictive modelling)
- **Loss function:** Quantifies the discrepancy between observed data and predictions

Linear regression - An old friend



- **Data:** Training data consists of independent pairs

$$(y_i, \mathbf{x}_i), \quad i = 1, \dots, n$$

Observed response $y_i \in \mathbb{R}$ for predictors $\mathbf{x}_i \in \mathbb{R}^p$ and design matrix $\mathbf{X}$ has rank $p + 1$

- **Model:**

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$ independent

- **Loss function:** **Least squares** solves standard linear regression problems, i.e. **squared error loss**

$$L(y, \hat{y}) = (y - \hat{y})^2 = \left(y - \mathbf{x}^T \underbrace{\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right)}_{=\hat{\boldsymbol{\beta}}}\right)^2$$

Statistical decision theory for regression (I)

- Squared error loss between outcome y and a prediction $f(\mathbf{x})$ dependent on the variable(s) x

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$$

- Assume we want to find the "best" f that can be learned from training data
- When a new pair of data $(y, \mathbf{x})$ from the same distribution (population) as the training data arrives, **expected prediction loss** for a given f is

$$J(f) = \mathbb{E}_{p(\mathbf{x},y)}\left[L(y, f(\mathbf{x}))\right] = \mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}_{p(y|\mathbf{x})}\left[L(y, f(\mathbf{x}))\right]\right]$$

- Define "best" by:

$$\widehat{f} = \arg\min_{f} J(f)$$

Statistical decision theory for regression (II)

- It can be derived (see blackboard) that

$$\hat{f}(\mathbf{x}) = \mathbb{E}_{p(y|\mathbf{x})}[y]$$

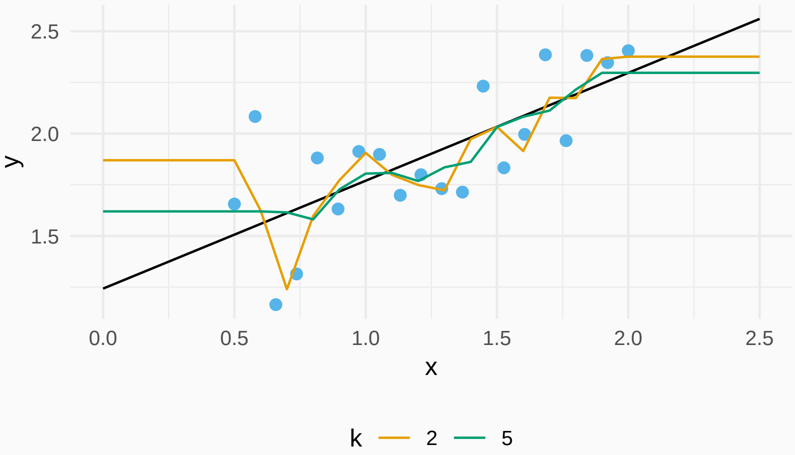
  the expectation of y given that $\mathbf{x}$ is fixed (conditional mean)
- Regression methods approximate the conditional mean
- For many observations y with identical $\mathbf{x}$ we could use

$$\mathbb{E}_{p(y|\mathbf{x})}[y] \approx \frac{1}{|\{y_i : \mathbf{x}_i = \mathbf{x}\}|} \sum_{\mathbf{x}_i = \mathbf{x}} y_i$$

- Probably more realistic to look for the k closest neighbours of $\mathbf{x}$ in the training data $N_k(\mathbf{x}) = \{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}\}$. Then

$$\mathbb{E}_{p(y|\mathbf{x})}[y] \approx \frac{1}{k} \sum_{\mathbf{x}_{i_l} \in N_k(\mathbf{x})} y_{i_l}$$

Average of k neighbours

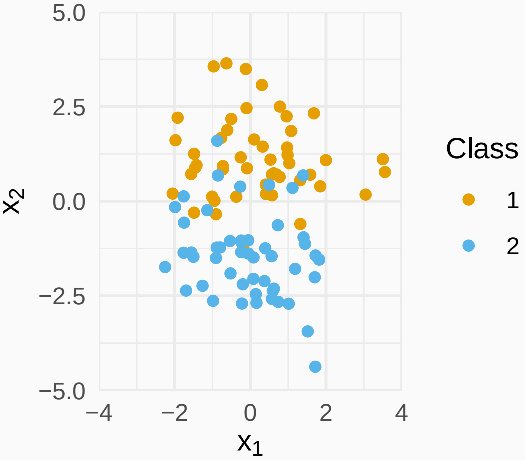


Back to linear regression

Linear regression is a **model-based approach** and assumes that the dependence of y on $\mathbf{x}$ can be written as a weighted sum

$$\mathbb{E}_{p(y|x)}[y] \approx \mathbf{x}^T \boldsymbol{\beta}$$

A simple example of classification



How do we classify a pair of new coordinates $\mathbf{x} = (x_1, x_2)$?

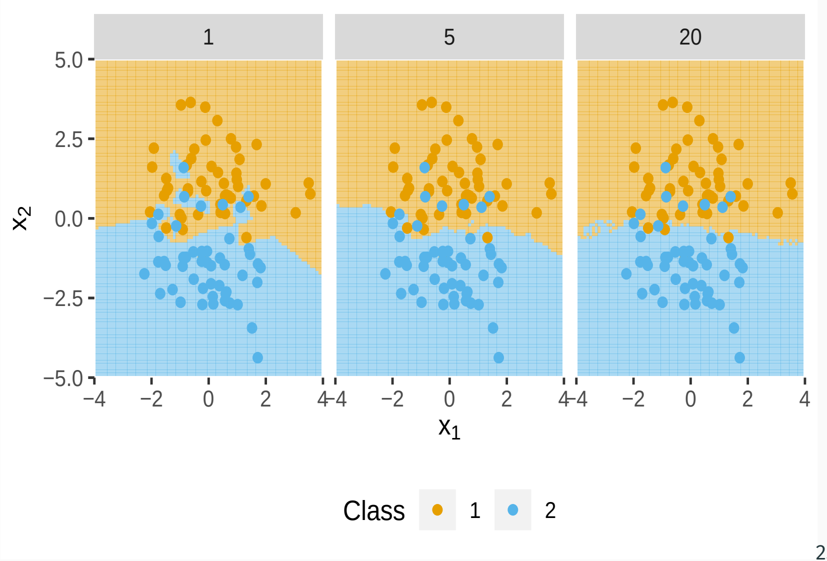
k-nearest neighbour classifier (kNN)

- Find the k predictors

$$N_k(\mathbf{x}) = \{\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}\}$$

  in the training sample, that are closest to $\mathbf{x}$ in the Euclidean norm.
- **Majority vote:** Assign $\mathbf{x}$ to the class that most predictors in $N_k(\mathbf{x})$ belong to (highest frequency)

kNN and its decision boundaries



Classification and Statistical Learning

Classification

Learn a rule $c(\mathbf{x})$ from data which maps observed features $\mathbf{x}$ to classes $\{1, \dots, K\}$.

Remember:

Statistical Learning

Learn a model from data by minimizing expected prediction error determined by a loss function.

Here: rule $\simeq$ model, and observed classes give us the required outcomes for learning.

What is a suitable loss?

Statistical decision theory for classification

- **0-1 misclassification loss:** Let i be the actual class of an object and $c(\mathbf{x})$ is a rule that returns the class for the variable(s) $\mathbf{x}$, then

$$L(i, c(\mathbf{x})) = \begin{cases} 0 & i = c(\mathbf{x}), \\ 1 & i \neq c(\mathbf{x}) \end{cases}$$

- As for regression, minimizing expected prediction error leads to the rule (see blackboard)

$$\hat{c}(\mathbf{x}) = \underset{1 \leq i \leq K}{\arg\max}\, p(i|\mathbf{x})$$

This is called **Bayes' rule**.

Back to kNN

- kNN solves the classification problem by approximating $p(i|\mathbf{x})$ with the frequency of class i among the k closest neighbours of $\mathbf{x}$.
- Given data $(i_l, \mathbf{x}_l)$ for $l = 1, \dots, n$ it holds that

$$\hat{c}(\mathbf{x}) = \underset{1 \le i \le K}{\arg\max} \frac{1}{k} \sum_{\mathbf{x}_l \in N_k(\mathbf{x})} \mathbb{1}(i_l = i)$$

Take-home message

- Big Data is complex and is multi-faceted
- Regression and classification can be formulated in the framework of Statistical Learning
- In both cases, focus is on prediction