Lecture 1: Introduction

Felix Held, Mathematical Sciences

MSA220/MVE440 Statistical Learning for Big Data

23rd March 2020

CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

What is Big Data?

Just a buzz word?

The cure for everything?

Business Insider[1]

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

Big Data - Big Problems?

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

Scientific discussion article[1]

[1] Lazer2014

Big Data - Big Problems?

Big data: are we making a big mistake?

Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media

Tim Harford MARCH 28, 2014

86

Financial Times[1]

The New York Times

Opinion

THE STONE

How Democracy Can Survive Big Data

By Colin Koopman

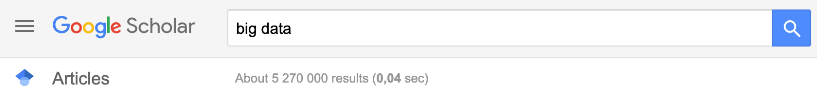
March 22, 2018

New York Times[2]

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

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

It’s a huge topic in science!



Over 5 million hits on Google Scholar

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*

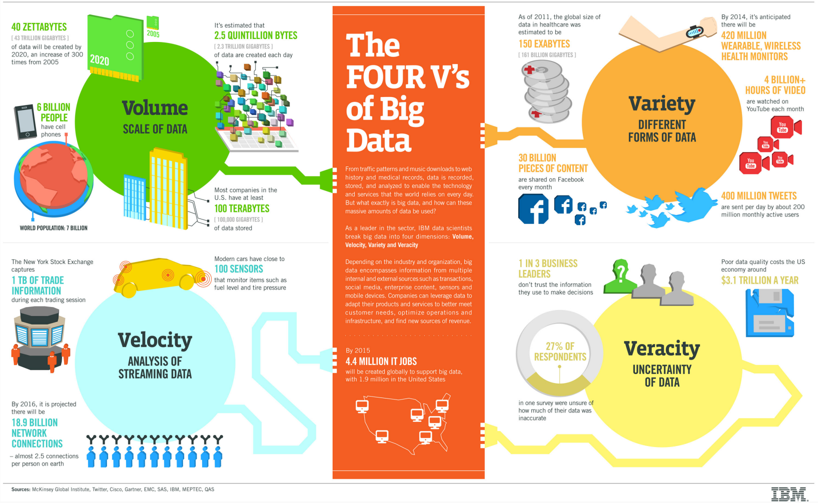
  Big-p setting

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

  Big-n / Big-p setting

Is this all?

The Four Vs of Big Data



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

How does statistics come into play?

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

- experimental design or ‘how to collect the data’
- 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

Statistical Learning

Basics about random variables

- We will consider **discrete** and **continuous** random quantities
- **Probability mass function (pmf)** $p(k)$ for a discrete variable
  Example: Bernoulli distribution with parameter $\theta \in (0, 1)$
  $$p(0) = \theta, \quad p(1) = 1 - \theta$$
- **Probability density function (pdf)** $p(\mathbf{x})$ for a continuous variables
  Example: Multivariate normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$
  $$p(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

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:

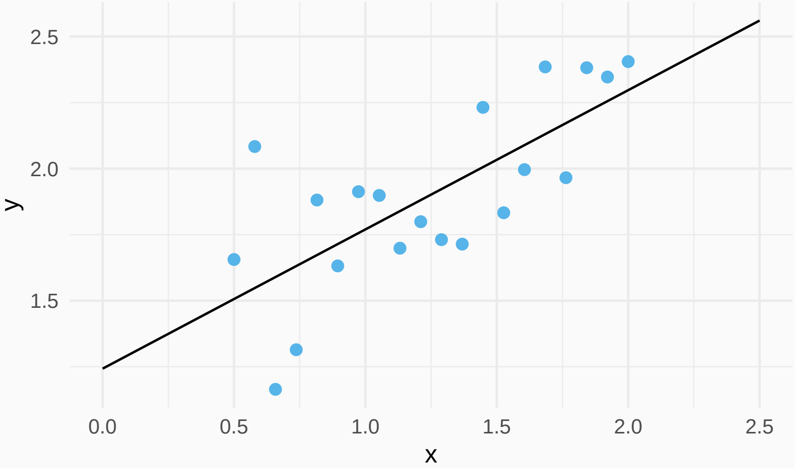
- $\mathbb{E}_{p(x)}[f(x)] = \int f(x)p(x)\,\mathrm{d}x$
- $\mathrm{Var}_{p(x)}[f(x)] = \mathbb{E}_{p(x)}\left[\left(f(x) - \mathbb{E}_{p(x)}[f(x)]\right)^2\right]$

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



Statistical Learning and Linear Regression

- **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 $\mathbf{x}_i \in \mathbb{R}^p$

- **Model:**

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

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

- **Loss function:** **Squared error loss**

$$L(y, \hat{y}) = (y - \hat{y})^2$$

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)

Can we determine $\widehat{f}$? Focus on inner expectation

$$\begin{aligned}
\mathbb{E}_{p(y|\mathbf{x})}\left[(y - f(\mathbf{x}))^2\right] &= \int (y - \mathbb{E}_{p(y|\mathbf{x})}[y] + \mathbb{E}_{p(y|\mathbf{x})}[y] - f(\mathbf{x}))^2 p(y|\mathbf{x})\,\mathrm{d}y \\
&= \int (y - \mathbb{E}_{p(y|\mathbf{x})}[y])^2 p(y|\mathbf{x})\,\mathrm{d}y \\
&+ 2\int (y - \mathbb{E}_{p(y|\mathbf{x})}[y])(\mathbb{E}_{p(y|\mathbf{x})}[y] - f(\mathbf{x}))p(y|\mathbf{x})\,\mathrm{d}y \\
&+ \int (\mathbb{E}_{p(y|\mathbf{x})}[y] - f(\mathbf{x}))^2 p(y|\mathbf{x})\,\mathrm{d}y \\
&= \mathrm{Var}_{p(y|\mathbf{x})}[y] + (\mathbb{E}_{p(y|\mathbf{x})}[y] - f(\mathbf{x}))^2
\end{aligned}$$

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

Statistical decision theory for regression (III)

- We just derived that

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

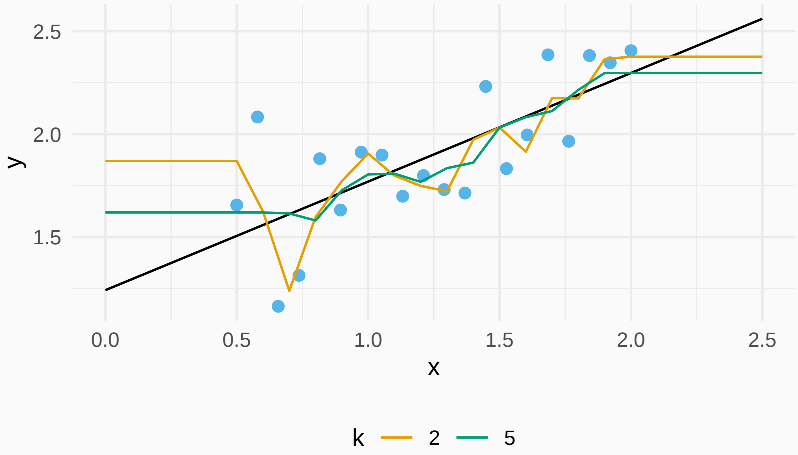
  the expectation of y given that $\mathbf{x}$ is fixed (conditional mean)
- Regression methods approximate the conditional mean
- For many observations y with identical $\mathbf{x}$ we could use

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

- 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, i.e.

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon$$

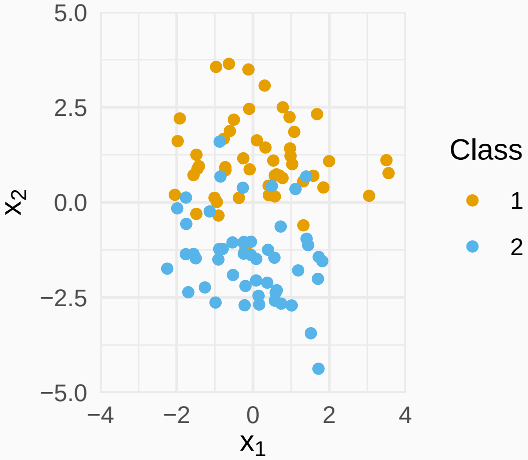
where $\varepsilon \sim N(0, 1)$. This implies to the mean of y given $\mathbf{x}$

$$\mathbb{E}_{p(y|x)}[y] = \mathbf{x}^\top \boldsymbol{\beta}.$$

Note that in practice this equality will only hold approximately.

Classification

A simple example of classification



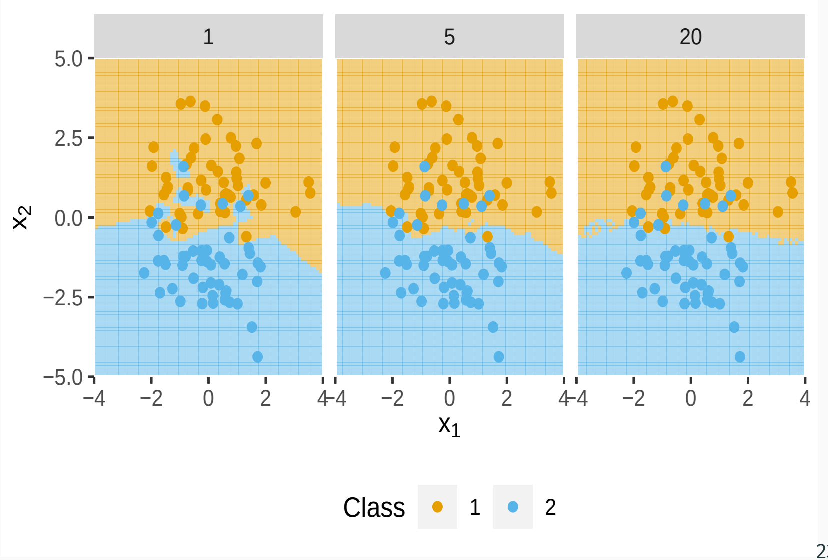
How do we classify a pair of new coordinates $\mathbf{x} = (x_1, x_2)$?

- Find the k predictors

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

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

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

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

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

- Expected prediction error

$$J(c) = \mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}_{p(i|\mathbf{x})}[\mathbb{1}(i \neq c(\mathbf{x}))]\right]$$

- Minimizing expected prediction error leads to the rule

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

This is called **Bayes' rule**.

Deriving Bayes' rule

Again, focus on inner expectation

$$\begin{aligned}\mathbb{E}_{p(i|\mathbf{x})}[\mathbb{1}(i \neq c(\mathbf{x}))] &= \sum_{i=1}^{K} \mathbb{1}(i \neq c(\mathbf{x}))p(i|\mathbf{x}) \\ &= \sum_{i \neq c(\mathbf{x})} p(i|\mathbf{x}) \\ &= 1 - p(c(\mathbf{x})|\mathbf{x})\end{aligned}$$

Minimal for $\hat{c}(\mathbf{x}) = \arg\max_{1 \leq i \leq K} p(i|\mathbf{x})$

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

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

A note on kNN

There are two choices to make when implementing a kNN method

1. The metric to determine a neighbourhood
   - e.g. Euclidean/ℓ_2 norm, Manhattan/ℓ_1 norm, max norm, …
2. The number of neighbours, i.e. k

The choice of metric changes the underlying local model of the method while k determines the size of this local model.

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