Lecture 1: Introduction

Rebecka Jörnsten, Mathematical Sciences

MSA220/MVE441 Statistical Learning for Big Data

24$^{\mathrm{nd}}$ March 2025

CHALMERS UNIVERSITY OF TECHNOLOGY | UNIVERSITY OF GOTHENBURG

The Data Deluge

- Data is getting cheap! (sometimes)
- Massive data collections across all fields! Molecular biology, health care, banking, streaming, marketing, citizen science, climate modeling, imaging, data feeds like twitter, ...
- What do we do with all this data?

The cure for everything is more data?

- **High-dimensional data: we obtain a more ”complete” picture of the world as opposed to having selected (perhaps because of technical constraints) to view a narrow perspective**

The cure for everything is more data?

- **High-dimensional data: we obtain a more "complete" picture of the world as opposed to having selected (perhaps because of technical constraints) to view a narrow perspective**
- but, if we collect tons of information about objects just because we can - do we risk drowning in uninformative data?

The cure for everything is more data?

- **High-dimensional data: we obtain a more ”complete” picture of the world as opposed to having selected (perhaps because of technical constraints) to view a narrow perspective**
- but, if we collect tons of information about objects just because we can - do we risk drowning in uninformative data?
- many methods perform poorly if we overwhelm them with high-dimensional data

The cure for everything is more data?

- **High-dimensional data: we obtain a more ”complete” picture of the world as opposed to having selected (perhaps because of technical constraints) to view a narrow perspective**
- but, if we collect tons of information about objects just because we can - do we risk drowning in uninformative data?
- many methods perform poorly if we overwhelm them with high-dimensional data
- is there a selection bias when we collect high-dimensional data? - which features of objects are easy to obtain?

The cure for everything is more data?

- **Big n data: we obtain a more "complete" picture of the population of patients/customers/units rather than a small sample which might not give us enough power to draw conclusions**

The cure for everything is more data?

- **Big n data: we obtain a more ”complete” picture of the population of patients/customers/units rather than a small sample which might not give us enough power to draw conclusions**
- Big n data can be associated with poor design - which type of observations are we collecting? selection bias, self reporting, unbalanced subpopulations in data, ...

The cure for everything is more data?

- **Big n data: we obtain a more ”complete” picture of the population of patients/customers/units rather than a small sample which might not give us enough power to draw conclusions**
- Big n data can be associated with poor design - which type of observations are we collecting? selection bias, self reporting, unbalanced subpopulations in data, ...
- Big n can lead to computational problems, can be more difficult to handle at the data exploration stage, difficult to visualize and assess model adequacy

The cure for everything is more data?

- **Big n data: we obtain a more ”complete” picture of the population of patients/customers/units rather than a small sample which might not give us enough power to draw conclusions**
- Big n data can be associated with poor design - which type of observations are we collecting? selection bias, self reporting, unbalanced subpopulations in data, ...
- Big n can lead to computational problems, can be more difficult to handle at the data exploration stage, difficult to visualize and assess model adequacy
- Big n statistics: with a big enough n everything becomes significant - think carefully about what the analysis goals are

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

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 ob
with entrepreneurs, scientists, governments and the media

Tim Harford MARCH 28, 2014

The New York Times

Opinion

THE STONE

How Democracy Can Survive Big Data

By Colin Koopman

March 22, 2018

Financial Times[1]

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!

Lot’s of research! Companies need the competence!

- Logistics, transport, banking, risk analysis, automated detection, image and video processing, molecular and systems biology
- and more...
- Methodology research: dimension reduction and visualization, computing solutions, feature selection/interpretability, scalable algorithms, ...

So Big Data is about size?

Yes and no.

So Big Data is about size?

Yes and no.

Note that *size* is a flexible term.

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

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

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

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*

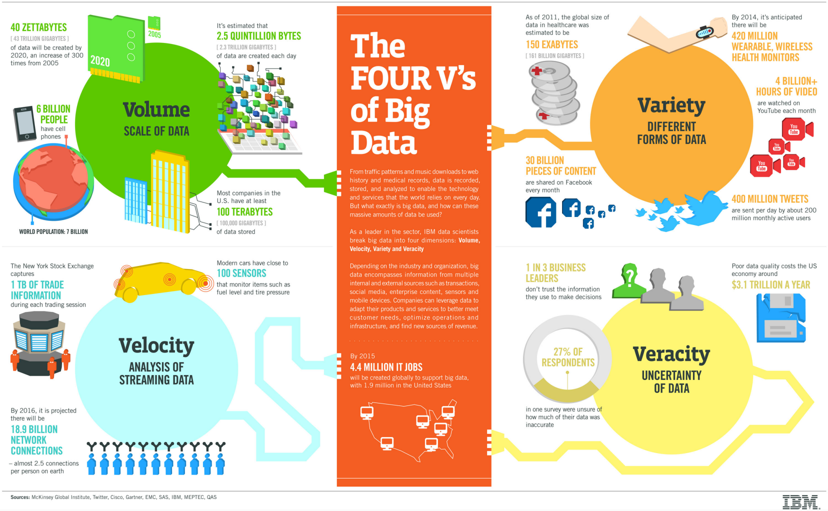
  Big-p setting

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

  Big-n / Big-p setting

Is this all?

The Four Vs of Big Data



http://web.archive.org/web/20210506042232/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’

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

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

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

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)

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

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

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

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

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 challenges in Big Data

- Proceed with caution: Big data does not mean that you can ignore assumptions, that large volumes of data ”solve everything”.

Statistical challenges in Big Data

- Proceed with caution: Big data does not mean that you can ignore assumptions, that large volumes of data ”solve everything”.
- Problem: exploration and visualization may be difficult/expensive but you *need* to check assumptions! Cannot rely on automation!

Statistical challenges in Big Data

- Proceed with caution: Big data does not mean that you can ignore assumptions, that large volumes of data ”solve everything”.
- Problem: exploration and visualization may be difficult/expensive but you *need* to check assumptions! Cannot rely on automation!
- In fact: **with Big Data you may encounter selection bias, mislabeled observations, imbalance of data, noisy data, spurious correlations, ...**

Statistical challenges in Big Data

- Proceed with caution: Big data does not mean that you can ignore assumptions, that large volumes of data ”solve everything”.
- Problem: exploration and visualization may be difficult/expensive but you *need* to check assumptions! Cannot rely on automation!
- In fact: **with Big Data you may encounter selection bias, mislabeled observations, imbalance of data, noisy data, spurious correlations, ...**
- Before analysis, spend time with the people collecting the data to try to understand how, why, think about sources for bias, questions that were *not* asked, ... **Taking time to understand the data will save you a lot of headache later on!**

Statistical challenges in Big Data

- Big p problems:

Statistical challenges in Big Data

- Big p problems:
  - Curse of dimensionality - notion of close/far breaks down

Statistical challenges in Big Data

- Big p problems:
  - Curse of dimensionality - notion of close/far breaks down
  - Multiple testing - spurious correlations

Statistical challenges in Big Data

- Big p problems:
  - Curse of dimensionality - notion of close/far breaks down
  - Multiple testing - spurious correlations
  - But also *Blessing of Dimensionality* - matrix completion, imputation - when there is implied structure in data.

Statistical challenges in Big Data

- Big p problems:
  - Curse of dimensionality - notion of close/far breaks down
  - Multiple testing - spurious correlations
  - But also *Blessing of Dimensionality* - matrix completion, imputation - when there is implied structure in data.
- Big n problems:

Statistical challenges in Big Data

- Big p problems:
  - Curse of dimensionality - notion of close/far breaks down
  - Multiple testing - spurious correlations
  - But also *Blessing of Dimensionality* - matrix completion, imputation - when there is implied structure in data.
- Big n problems:
  - Selection bias
  - Significance vs Importance

Statistical challenges in Big Data

- Big p problems:
  - Curse of dimensionality - notion of close/far breaks down
  - Multiple testing - spurious correlations
  - But also *Blessing of Dimensionality* - matrix completion, imputation - when there is implied structure in data.
- Big n problems:
  - Selection bias
  - Significance vs Importance
  - Overfitting

Statistical challenges in Big Data

- Big p problems:
  - Curse of dimensionality - notion of close/far breaks down
  - Multiple testing - spurious correlations
  - But also *Blessing of Dimensionality* - matrix completion, imputation - when there is implied structure in data.
- Big n problems:
  - Selection bias
  - Significance vs Importance
  - Overfitting
  - Computational burden

Statistical challenges in Big Data

- Big p problems:
  - Curse of dimensionality - notion of close/far breaks down
  - Multiple testing - spurious correlations
  - But also *Blessing of Dimensionality* - matrix completion, imputation - when there is implied structure in data.
- Big n problems:
  - Selection bias
  - Significance vs Importance
  - Overfitting
  - Computational burden
  - but also blessing (e.g. SGD and ensembles, ”data hungry” methods like NN, rare-events detection, noise correction)

PAUSE!!!!

A bit of recap.....

If what comes next feels unfamiliar - spend a bit of time reviewing basic stat concepts

Terminology, Background and Basics

1. Statistics terminology recap
2. Basics of statistical learning
3. Setting the stage for the course

Basics about random variables

- We will consider **discrete** and **continuous** random quantities

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 $\theta \in (0, 1)$

$$p(0) = \theta, \quad p(1) = 1 - \theta$$

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 $\theta \in (0, 1)$

$$p(0) = \theta, \quad p(1) = 1 - \theta$$

- Where will we see this? **Classification**, where θ denotes the probability of class 1 and can be observation specific, θ_i, and depend on features of this observation, $\theta(x_i)$

Basics about random variables

- We will consider **discrete** and **continuous** random quantities

- Where will we see this? **Classification and clustering**, describing the dependency between features. The distribution parameters may be class specific μ_k, Σ_k, class $k \in \{1, \cdots, K\}$

Basics about random variables

- We will consider **discrete** and **continuous** random quantities
- **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 $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$

$$p(\mathbf{x}) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- Where will we see this? **Classification and clustering**, describing the dependency between features. The distribution parameters may be class specific μ_k, Σ_k, class $k \in \{1, \cdots, K\}$

Two important rules (and a consequence)

Marginalisation

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

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

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

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

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**

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

Two important rules (and a consequence)

- Where will we see this?

Two important rules (and a consequence)

- Where will we see this?
- Averaging out (marginalization) over training and test data to quantify performance of prediction methods, understanding uncertainties of estimation

Two important rules (and a consequence)

- Where will we see this?
- Averaging out (marginalization) over training and test data to quantify performance of prediction methods, understanding uncertainties of estimation
- Conditioning in classification/regression where we treat features of objects as known and focus on the random variation of the outcome (response, class) only, *given* the features

Two important rules (and a consequence)

- Where will we see this?
- Averaging out (marginalization) over training and test data to quantify performance of prediction methods, understanding uncertainties of estimation
- Conditioning in classification/regression where we treat features of objects as known and focus on the random variation of the outcome (response, class) only, *given* the features
- Bayes law: the basis for building classification rules by updating marginal population frequencies once we observe observation specific features.

Expectation and variance

Expectations and variance depend on an underlying pdf/pmf.

Notation:

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

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

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

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

What is Statistical Learning?

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

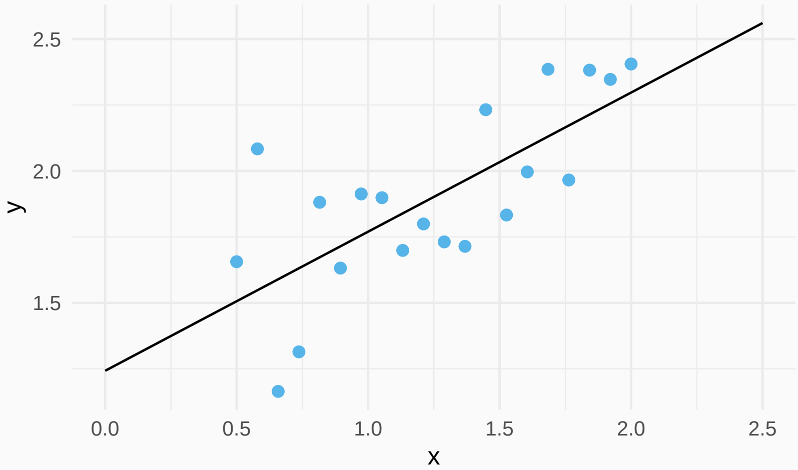
- **Model:** Find a model that is suitable for the data
- **Data:** Data with known outcomes is needed
- **Expected prediction error:** Focus on quality of prediction (predictive modelling)

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

- **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

- **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**

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

PAUSE 2!!!!!

- The 5 basic assumptions in linear regression are....?

PAUSE 2!!!!!

- The 5 basic assumptions in linear regression are....?
- What can happen if these are violated?

PAUSE 2!!!!!

- The 5 basic assumptions in linear regression are....?
- What can happen if these are violated?
- What, if anything, can we do to handle violations?

Statistical decision theory for regression (I)

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

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

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

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

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

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 $\widehat{f}$?

Statistical decision theory for regression (II)

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

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

Statistical decision theory for regression (II)

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

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

$\mathbb{E}_{p(y|\mathbf{x})}[y]$: "Average of y in vertical slice defined by x"
f(x): Our model

Statistical decision theory for regression (II)

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

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

Statistical decision theory for regression (II)

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

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

Statistical decision theory for regression (II)

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

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

Statistical decision theory for regression (II)

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

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

Statistical decision theory for regression (II)

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

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

Statistical decision theory for regression (II)

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

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

Statistical decision theory for regression (II)

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

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

Statistical decision theory for regression (II)

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

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

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)

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

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

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

Statistical decision theory for regression (III)

- We just derived that

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

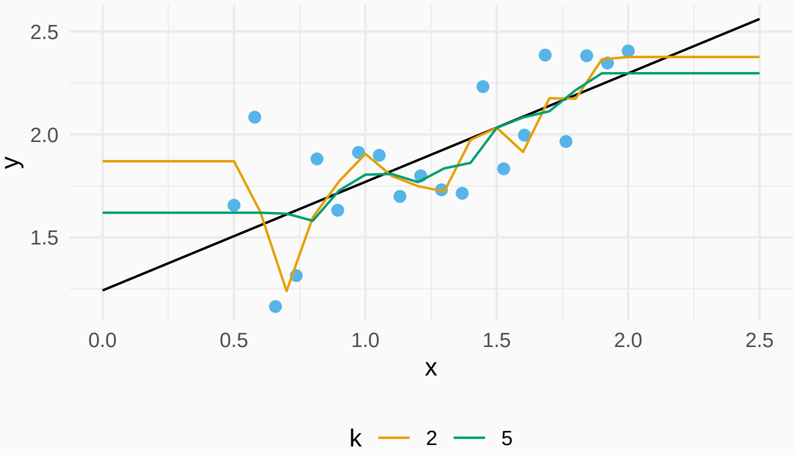
  the expectation of y given that $\mathbf{x}$ is fixed (conditional mean)
- Regression methods approximate the conditional mean
- For many observations y with identical $\mathbf{x}$ we could use

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

- 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, i.e.

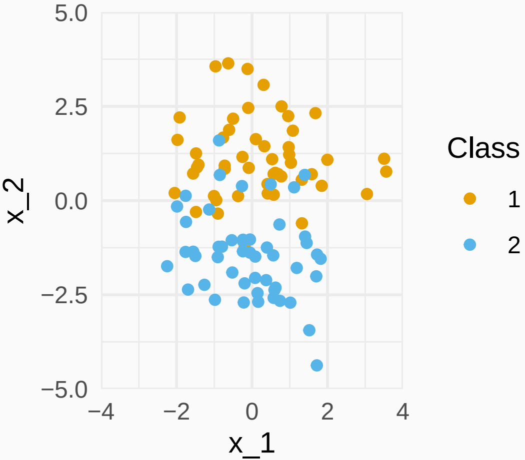
$$y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$. The mean of y given $\mathbf{x}$ is therefore

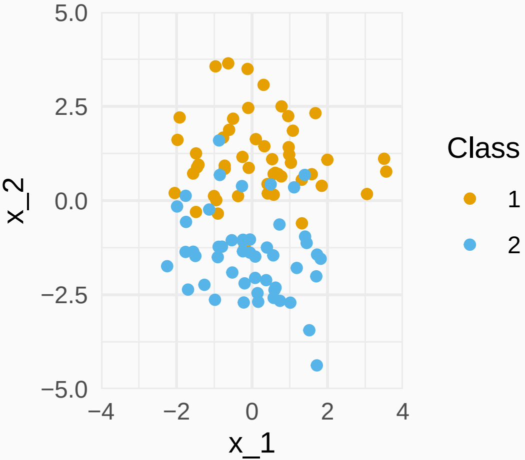
$$\mathbb{E}_{p(y|x)}[y] = \mathbf{x}^\top \boldsymbol{\beta}.$$

Note that in practice this equality will only hold approximately.

A simple example of 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

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

in the training sample, that are closest to $\mathbf{x}$ in the Euclidean norm.

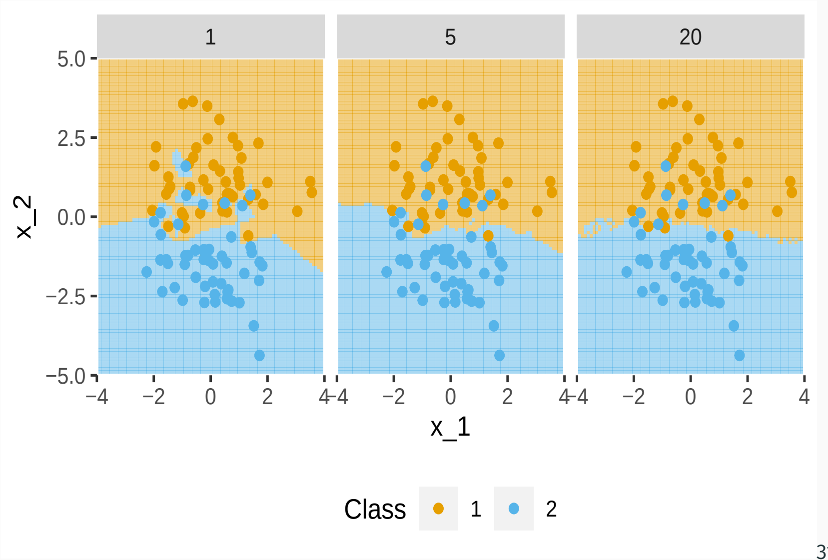
k-nearest neighbour classifier (kNN)

- 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

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

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.

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.

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

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

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

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

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

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

Deriving Bayes' rule

Again, focus on inner expectation

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

Deriving Bayes' rule

Again, focus on inner expectation

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

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

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

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.

Summary

- Prediction rules or models are all about capturing the *conditional expectation* of our outcome y, $\mathbb{E}_{p(y|\mathbf{x})}[y]$, given x (our features, variables)

Summary

- Prediction rules or models are all about capturing the *conditional expectation* of our outcome y, $\mathbb{E}_{p(y|\mathbf{x})}[y]$, given x (our features, variables)
  - MSE loss → conditional mean

Summary

- Prediction rules or models are all about capturing the *conditional expectation* of our outcome y, $\mathbb{E}_{p(y|\mathbf{x})}[y]$, given x (our features, variables)
  - MSE loss $\rightarrow$ conditional mean
  - 0-1 loss/classification $\rightarrow$ conditional probability, majority rule

Summary

- Prediction rules or models are all about capturing the *conditional expectation* of our outcome y, $\mathbb{E}_{p(y|\mathbf{x})}[y]$, given x (our features, variables)
  - MSE loss → conditional mean
  - 0-1 loss/classification → conditional probability, majority rule
- Some other methods:

Summary

- Prediction rules or models are all about capturing the *conditional expectation* of our outcome y, $\mathbb{E}_{p(y|\mathbf{x})}[y]$, given x (our features, variables)
  - MSE loss → conditional mean
  - 0-1 loss/classification → conditional probability, majority rule
- Some other methods:
  - CART, Random Forest: assuming locally constant conditional means

Summary

- Prediction rules or models are all about capturing the *conditional expectation* of our outcome y, $\mathbb{E}_{p(y|\mathbf{x})}[y]$, given x (our features, variables)
  - MSE loss → conditional mean
  - 0-1 loss/classification → conditional probability, majority rule
- Some other methods:
  - CART, Random Forest: assuming locally constant conditional means
  - Logistic regression: parameterize the conditional mean, logits. Discriminant analysis: model both x and y, not just conditional

Summary

- Prediction rules or models are all about capturing the *conditional expectation* of our outcome y, $\mathbb{E}_{p(y|\mathbf{x})}[y]$, given x (our features, variables)
  - MSE loss → conditional mean
  - 0-1 loss/classification → conditional probability, majority rule
- Some other methods:
  - CART, Random Forest: assuming locally constant conditional means
  - Logistic regression: parameterize the conditional mean, logits. Discriminant analysis: model both x and y, not just conditional
  - Flexibility of feature space: how we compute distance in kNN, transformation of features, ”the kernel trick” (e.g. used in svms, hinge loss)

Summary

- Prediction rules or models are all about capturing the *conditional expectation* of our outcome y, $\mathbb{E}_{p(y|\mathbf{x})}[y]$, given x (our features, variables)
  - MSE loss → conditional mean
  - 0-1 loss/classification → conditional probability, majority rule
- Some other methods:
  - CART, Random Forest: assuming locally constant conditional means
  - Logistic regression: parameterize the conditional mean, logits. Discriminant analysis: model both x and y, not just conditional
  - Flexibility of feature space: how we compute distance in kNN, transformation of features, "the kernel trick" (e.g. used in svms, hinge loss)
  - NN: generate flexible features with the first layers, utilize with the last layer as in standard methods

Summary

- Prediction rules or models are all about capturing the *conditional expectation* of our outcome y, $\mathbb{E}_{p(y|\mathbf{x})}[y]$, given x (our features, variables)
  - MSE loss → conditional mean
  - 0-1 loss/classification → conditional probability, majority rule
- Some other methods:
  - CART, Random Forest: assuming locally constant conditional means
  - Logistic regression: parameterize the conditional mean, logits. Discriminant analysis: model both x and y, not just conditional
  - Flexibility of feature space: how we compute distance in kNN, transformation of features, "the kernel trick" (e.g. used in svms, hinge loss)
  - NN: generate flexible features with the first layers, utilize with the last layer as in standard methods

Questions

- Drawback/danger with L2/MSE loss? Drawback/danger with 0-1 loss?

Questions

- Drawback/danger with L2/MSE loss? Drawback/danger with 0-1 loss?
- What if features have different distributions - what does this mean for kNN?

Questions

- Drawback/danger with L2/MSE loss? Drawback/danger with 0-1 loss?
- What if features have different distributions - what does this mean for kNN?
- Big N problems in regression? What about kNN?

Questions

- Drawback/danger with L2/MSE loss? Drawback/danger with 0-1 loss?
- What if features have different distributions - what does this mean for kNN?
- Big N problems in regression? What about kNN?
- Big p problems in regression? What about kNN?

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
- Next class: Dimension reduction, Logistic regression and Bias-Variance trade-off