Introduction to Data Science - 1MS041

Benny Avelin

Department of Mathematics

HT 2022

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Specifics

1. The hand in part consists of 3 group assignments.

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Specifics

1. The hand in part consists of 3 group assignments.
2. The groups are free to form in the size-range 3-5 ppl.

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Specifics

1. The hand in part consists of 3 group assignments.
2. The groups are free to form in the size-range 3-5 ppl.
3. Individual contribution is mandatory.

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Specifics

1. The hand in part consists of 3 group assignments.
2. The groups are free to form in the size-range 3-5 ppl.
3. Individual contribution is mandatory.
4. To pass you need to have completed all 3 group assignments.

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Specifics

1. To pass the exam you need 20p/40p or higher, including bonus from individual assignments.

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Specifics

1. To pass the exam you need 20p/40p or higher, including bonus from individual assignments.
2. We have 4 computer based individual assignments which are automatically graded, each consists of 24p.

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Specifics

1. To pass the exam you need 20p/40p or higher, including bonus from individual assignments.
2. We have 4 computer based individual assignments which are automatically graded, each consists of 24p.
3. These give bonus points for the final exam.

Introduction to Data Science

Base outline

1. 2.5hp Hand in assignments (3 group assignments)
2. 7.5hp Final exam in January

Specifics

1. To pass the exam you need 20p/40p or higher, including bonus from individual assignments.
2. We have 4 computer based individual assignments which are automatically graded, each consists of 24p.
3. These give bonus points for the final exam.
4. The bonus for the main exam is the score for the individual assignment / 16. So if you get full score, i.e. 96p then you have a bonus of 6p on the final exam (a pretty good bonus).

Practical information

1. The final exam is computer based and will be in one of the computer labs in Ångström, and is similar to the computer based individual assignments.

Practical information

1. The final exam is computer based and will be in one of the computer labs in Ångström, and is similar to the computer based individual assignments.
2. On Friday we will have a lab walkthrough.
   - 2.1 We will be looking at how do practially do the computer exercises on the lab computers.
   - 2.2 The computer labs are fairly small, roughly 30 seats. Therefore you will be split into two labs.

Practical information

1. The final exam is computer based and will be in one of the computer labs in Ångström, and is similar to the computer based individual assignments.
2. On Friday we will have a lab walkthrough.
   - 2.1 We will be looking at how do practially do the computer exercises on the lab computers.
   - 2.2 The computer labs are fairly small, roughly 30 seats. Therefore you will be split into two labs.
3. For the group assignments, you are free to form your own groups. This should preferably be done before the end of the problem session on Thursday next week. Rules etc. can be found on the course website. Aim for 5 people in each group.

What we will learn

1. The goal is to provide a rigorous basis for your future studies in the program.

What we will learn

1. The goal is to provide a rigorous basis for your future studies in the program.
2. We will be mixing theoretical and practical during each lecture / problem session.

What we will learn

1. The goal is to provide a rigorous basis for your future studies in the program.
2. We will be mixing theoretical and practical during each lecture / problem session.
3. At the end of this course, you will be able to solve a few typical problems. But more importantly, you will have learned the language of probability which will allow you to understand other problems.

What you will learn: specifically

1. Basic probability
   1.1 Probability, Random Variables and Concentration.

What you will learn: specifically

1. Basic probability
   - 1.1 Probability, Random Variables and Concentration.
2. Simulation and Markov chains
   - 2.1 How to generate random numbers on the computer
   - 2.2 How to produce samples from a specific distribution
   - 2.3 How to model sequential problems using Markov chains

What you will learn: specifically

1. Basic probability
   - 1.1 Probability, Random Variables and Concentration.
2. Simulation and Markov chains
   - 2.1 How to generate random numbers on the computer
   - 2.2 How to produce samples from a specific distribution
   - 2.3 How to model sequential problems using Markov chains
3. Supervised learning
   - 3.1 How to model certain typical Supervised learning problems.
   - 3.2 Estimation
   - 3.3 Pattern Recognition (Classification) and Regression
   - 3.4 Validation with rigorous guarantees.

What you will learn: specifically

1. Basic probability
   - 1.1 Probability, Random Variables and Concentration.
2. Simulation and Markov chains
   - 2.1 How to generate random numbers on the computer
   - 2.2 How to produce samples from a specific distribution
   - 2.3 How to model sequential problems using Markov chains
3. Supervised learning
   - 3.1 How to model certain typical Supervised learning problems.
   - 3.2 Estimation
   - 3.3 Pattern Recognition (Classification) and Regression
   - 3.4 Validation with rigorous guarantees.
4. Unsupervised learning
   - 4.1 Dimensionality reduction
   - 4.2 Some anomaly detection

Model problem: SMS spam filtering

You would like to filter your SMS (or similar instant messaging) texts as "Spam" or not. Consider the following examples

1. *Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's*
2. *Nah I don't think he goes to usf, he lives around here though*

Model problem: SMS spam filtering

You would like to filter your SMS (or similar instant messaging) texts as "Spam" or not. Consider the following examples

1. *Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's*
2. *Nah I don't think he goes to usf, he lives around here though*

What about the following?

1. *Thanks for your subscription to Ringtone UK your mobile will be charged 5/month Please confirm by replying YES or NO. If you reply NO you will not be charged*

Model problem: SMS spam filtering (Estimation / Classification)

We can ask a few different questions:

1. What is the probability that an incoming text is spam?

Model problem: SMS spam filtering (Estimation / Classification)

We can ask a few different questions:

1. What is the probability that an incoming text is spam?
2. Given that you see the word "free" in the text, what is now the probability that it is spam?

Model problem: SMS spam filtering (Estimation / Classification)

We can ask a few different questions:

1. What is the probability that an incoming text is spam?
2. Given that you see the word "free" in the text, what is now the probability that it is spam?
3. Can we find certain indicator words that increase the probability that a text is spam?

Model problem: SMS spam filtering (Estimation / Classification)

We can ask a few different questions:

1. What is the probability that an incoming text is spam?
2. Given that you see the word "free" in the text, what is now the probability that it is spam?
3. Can we find certain indicator words that increase the probability that a text is spam?

What you will learn

1. What do we actually mean when we ask these questions. (Probability)
2. How to process data and actually solve them using the computer. (ETL)
3. How to test the found solution with guarantees. (Concentration)

Corporate travel (Markov chains)

Consider having access to source-destination of multiple business flight travelers.

User	Source	Destination
1	Aracaju (SE)	Recife (PE)
2	Florianopolis (SC)	Brasilia (DF)

Corporate travel (Markov chains)

Consider having access to source-destination of multiple business flight travelers.

User	Source	Destination
1	Aracaju (SE)	Recife (PE)
2	Florianopolis (SC)	Brasilia (DF)

We could here ask a few questions:

1. What is the probability that someone taking their first flight from Aracaju to Recife would end up in Brasilia? This could be used to recommend hotels.

Corporate travel (Markov chains)

Consider having access to source-destination of multiple business flight travelers.

User	Source	Destination
1	Aracaju (SE)	Recife (PE)
2	Florianopolis (SC)	Brasilia (DF)

We could here ask a few questions:

1. What is the probability that someone taking their first flight from Aracaju to Recife would end up in Brasilia? This could be used to recommend hotels.
2. What is the average cost going from Aracaju to Recife?

Corporate travel (Markov chains)

Consider having access to source-destination of multiple business flight travelers.

User	Source	Destination
1	Aracaju (SE)	Recife (PE)
2	Florianopolis (SC)	Brasilia (DF)

We could here ask a few questions:

1. What is the probability that someone taking their first flight from Aracaju to Recife would end up in Brasilia? This could be used to recommend hotels.
2. What is the average cost going from Aracaju to Recife?
3. What is the probability that a user starting in Recife will get back to Recife within 4 steps?

Corporate travel (Markov chains)

Consider having access to source-destination of multiple business flight travelers.

User	Source	Destination
1	Aracaju (SE)	Recife (PE)
2	Florianopolis (SC)	Brasilia (DF)

We could here ask a few questions:

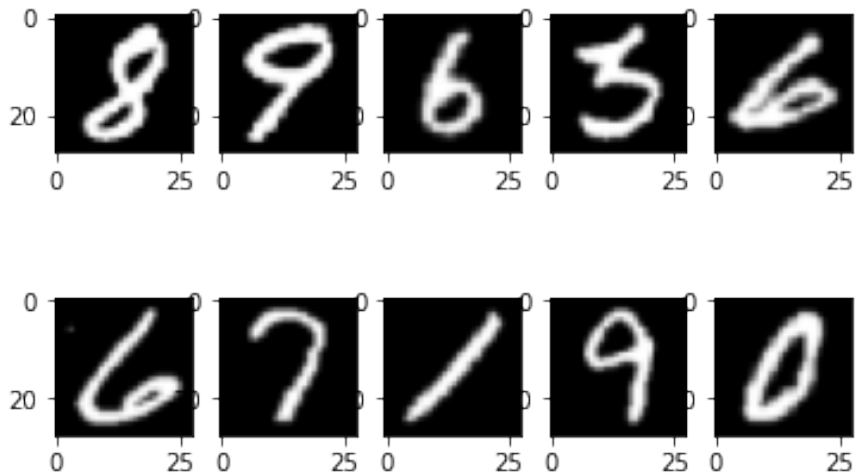
1. What is the probability that someone taking their first flight from Aracaju to Recife would end up in Brasilia? This could be used to recommend hotels.
2. What is the average cost going from Aracaju to Recife?
3. What is the probability that a user starting in Recife will get back to Recife within 4 steps?
4. Can we simulate our own flights?

Anomaly detection

- Anomaly detection, is a method one uses to distinguish something out of the ordinary. Usually this begins by us having a way to represent what is normal, and then check when something cannot be well represented anymore.
- This is used everywhere, from credit card fraud to predictive maintenance.

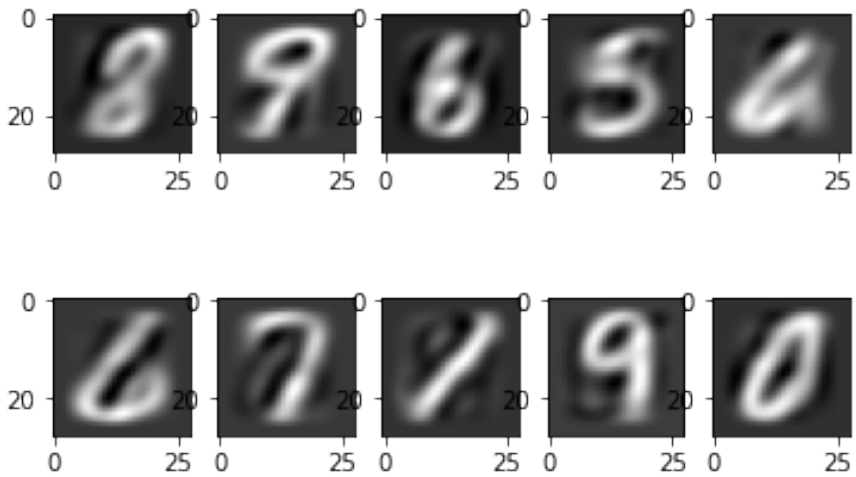
Dimensionality reduction for anomaly detection

Consider the following images of handwritten digits:

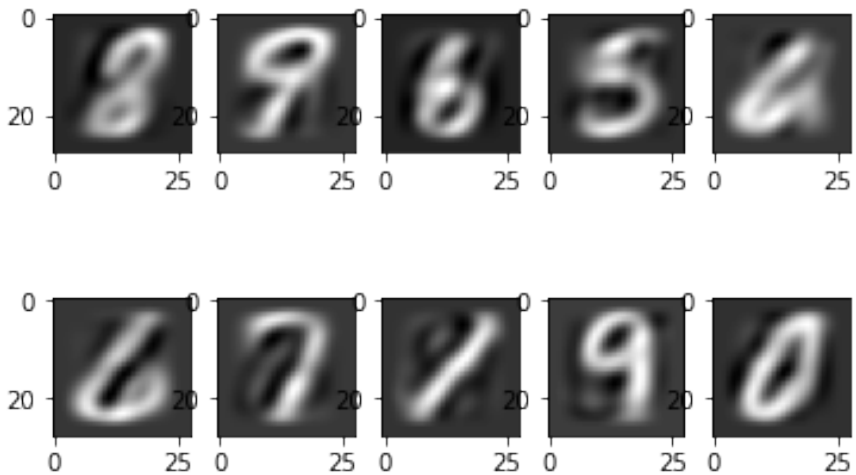


Can we represent these with less data?

Here I have used a technique called SVD to represent the images



Here I have used a technique called SVD to represent the images



But watch what happens when I supply something else