COMP229: Introduction to Data Science
Lecture 1: overview and expectations

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Background and contact details

2008 PhD Dynamical Systems, MSU.

till 2008: Lecturer at Higher School of Economics, Moscow.

2009-16 Teaching at Maths dept. in Durham Uni.

2017-19: Teaching in Computer Science.

2020-23: Data Scientist in Population Health, collaborating with Materials Innovation Factory.

2023: Computer Science.

Best contact: email O.Anosova@liverpool.ac.uk.

For all admin: e-mail csstudy@liverpool.ac.uk

What is important for COMP229?

The keyword is *Science*, which is "a systematic enterprise that builds and organises knowledge in the form of *testable explanations*" [Wikipedia].

Hence COMP229 will involve many *rigorous definitions* and *logical proofs*.

What is Data Science?

Definition 1.1. Data Science is an interdisciplinary field that uses *scientific methods* and algorithms to extract knowledge from data [Wikipedia].

Other less strict names for Data Science:
pattern recognition (old), data mining (new).

What is Data Science?

Definition 1.1. Data Science is an interdisciplinary field that uses *scientific methods* and algorithms to extract knowledge from data [Wikipedia].

Other less strict names for Data Science: pattern recognition (old), data mining (new).

Since many companies use their own software (or even languages), COMP229 will focus only on scientific foundations (not specific implementations) to prepare you for any good job on data analysis.

What's the problem with data?

What's the problem with data?

It's big! Data deluge / information explosion challenges:

What's the problem with data?

It's big! Data deluge / information explosion challenges:

Accuracy problems: data from untrusted sources, repetitions, trust in the measurements of accuracy.

What's the problem with data?

It's big! Data deluge / information explosion challenges:

Accuracy problems: data from untrusted sources, repetitions, trust in the measurements of accuracy.

Hard to identify patterns due to
curse of dimensionality, combinatorial explosion, higher false discovery rate with higher complexity.

What's the problem with data?

It's big! Data deluge / information explosion challenges:

Accuracy problems: data from untrusted sources, repetitions, trust in the measurements of accuracy.

Hard to identify patterns due to
curse of dimensionality, combinatorial explosion, higher false discovery rate with higher complexity.

Anonymosity becomes impossible.

What's the problem with data?

It's big! Data deluge / information explosion challenges:

Accuracy problems: data from untrusted sources, repetitions, trust in the measurements of accuracy.

Hard to identify patterns due to
curse of dimensionality, combinatorial explosion, higher false discovery rate with higher complexity.

Anonymosity becomes impossible.

Accessibility and cost!

Data analysis approaches

Most time is spent on data wrangling (collecting, cleaning, transforming), not analysis.

Statistics looks for underlying reasons of data.

Machine learning predicts future from given data.

Data analysis approaches

Most time is spent on data wrangling (collecting, cleaning, transforming), not analysis.

Statistics looks for underlying reasons of data.

Machine learning predicts future from given data.

Automated logic example:

- There is nothing better than eternal happiness.

Data analysis approaches

Most time is spent on data wrangling (collecting, cleaning, transforming), not analysis.

Statistics looks for underlying reasons of data.

Machine learning predicts future from given data.

Automated logic example:

- There is nothing better than eternal happiness.

- But a ham sandwich is better than nothing.

Data analysis approaches

Most time is spent on data wrangling (collecting, cleaning, transforming), not analysis.

Statistics looks for underlying reasons of data.

Machine learning predicts future from given data.

Automated logic example:

- There is nothing better than eternal happiness.

- But a ham sandwich is better than nothing.

- Hence, a ham sandwich is better that eternal happiness.

More examples of use of patterns

Olle Häggström's (Swedish statistician) example: future superAI has a task of increasing the overall happiness in the humanity, brilliantly solves it by computing that:

More examples of use of patterns

Olle Häggström's (Swedish statistician) example: future superAI has a task of increasing the overall happiness in the humanity, brilliantly solves it by computing that:
a) with current definitions the total overall happiness is negative, and

More examples of use of patterns

Olle Häggström's (Swedish statistician) example: future superAI has a task of increasing the overall happiness in the humanity, brilliantly solves it by computing that:
a) with current definitions the total overall happiness is negative, and
b) 0 is surely an increase from a negative number.

Is it all just stupid machines?

Human pattern recognition

A 2014 study showed a group of 20 individuals random greyscale photos.

They were told that 50% of photos contained faces.

The group reported seeing a face in

Human pattern recognition

A 2014 study showed a group of 20 individuals random greyscale photos.
They were told that 50% of photos contained faces.
The group reported seeing a face in 38% of the cases. But all the photos were just random, they were spotting a pattern where there wasn't one.

Data Science aims to find conditions on data when suitable algorithms guarantee correct predictions.

Human pattern recognition

A 2014 study showed a group of 20 individuals random greyscale photos.
They were told that 50% of photos contained faces.
The group reported seeing a face in 38% of the cases. But all the photos were just random, they were spotting a pattern where there wasn't one.

Data Science aims to find conditions on data when suitable algorithms guarantee correct predictions.

COMP229 will mainly discuss unsupervised learning.

Yann LeCun’s classification

Types of machine learning

Yann Lecun's Black Forest cake

- **"Pure" Reinforcement Learning** (cherry)
  - The machine predicts a scalar reward given once in a while.
  - **A few bits for some samples**
- **Supervised Learning** (icing)
  - The machine predicts a category or a few numbers for each input
  - Predicting human-supplied data
  - **10→10,000 bits per sample**
- **Unsupervised/Predictive Learning** (cake)
  - The machine predicts any part of its input for any observed part.
  - Predicts future frames in videos
  - **Millions of bits per sample**

- (Yes, I know, this picture is slightly offensive to RL folks. But I'll make it up)

Brief syllabus of COMP229

Fundamentals of statistics and probability theory

Brief syllabus of COMP229

Fundamentals of statistics and probability theory

Statistical learning: regression, hypotheses

Brief syllabus of COMP229

Fundamentals of statistics and probability theory

Statistical learning: regression, hypotheses

Equivalence relations and metrics on data, geometric invariants

Brief syllabus of COMP229

Fundamentals of statistics and probability theory

Statistical learning: regression, hypotheses

Equivalence relations and metrics on data,
geometric invariants

Clustering methods

Brief syllabus of COMP229

Fundamentals of statistics and probability theory

Statistical learning: regression, hypotheses

Equivalence relations and metrics on data, geometric invariants

Clustering methods

Linear algebra for dimensionality reduction

Brief syllabus of COMP229

Fundamentals of statistics and probability theory

Statistical learning: regression, hypotheses

Equivalence relations and metrics on data, geometric invariants

Clustering methods

Linear algebra for dimensionality reduction

Principal Component Analysis and SVD

Brief syllabus of COMP229

Fundamentals of statistics and probability theory

Statistical learning: regression, hypotheses

Equivalence relations and metrics on data, geometric invariants

Clustering methods

Linear algebra for dimensionality reduction

Principal Component Analysis and SVD

Do you have skills for COMP229? Probably, if you can compute this without a calculator and in one line: $51 \times 24 - 49 \times 93 + 27 \times 51 + 44 \times 49 = ?$

Lectures, tutorials and rooms

30 lectures in weeks 1-10, 10 tutorials in weeks 2-11.

Lectures + tutorials: 30+10 hours. Expected total: 150 hours. Can you guess where those extra work hours should come from?

Lectures, tutorials and rooms

30 lectures in weeks 1-10, 10 tutorials in weeks 2-11.

Lectures + tutorials: 30+10 hours. Expected total: 150 hours. Can you guess where those extra work hours should come from?

There's plenty of time for private study: attempting problems, going over notes, discussing with peers, reading around the subject.

Lectures, tutorials and rooms

30 lectures in weeks 1-10, 10 tutorials in weeks 2-11.

Lectures + tutorials: 30+10 hours. Expected total: 150 hours. Can you guess where those extra work hours should come from?

There's plenty of time for private study: attempting problems, going over notes, discussing with peers, reading around the subject.

Lectures: Monday at 14.00-15.00 and Friday 9.00 - 11.00 in MATH-029 (Forsyth Lecture Theatre).

Textbooks on Data Science

The textbooks below are in the reading list on CANVAS aren't strictly needed, only for enthusiasts.

- Applied Linear Algebra by *Olvert, Shakiban*
- Introduction to Statistics by *David Lane*
- Principles of Data Science by *Sinan Ozdemir*
- Learning from Data by *Glenberg, Andrzejewski*

If you find anything better, please e-mail me.

Collaboration

COMP229 is lectured in parallel to Computer Science and Maths students.

We will be taking a mathematical approach to data science...

CS students - be prepared for abstract definitions and rigorous proofs

Collaboration

COMP229 is lectured in parallel to Computer Science and Maths students.

We will be taking a mathematical approach to data science. . .

CS students - be prepared for abstract definitions and rigorous proofs

. . . but with practical goals in mind.

Maths students - be prepared for some 'top-level' thinking

Please collaborate and share your knowledge!

Resources on CANVAS

Slides will be regularly updated. If you notice typos, please e-mail me.

The discussion board is for asking constructive questions and answering them!

Resources on CANVAS

Slides will be regularly updated. If you notice typos, please e-mail me.

The discussion board is for asking constructive questions and answering them!

Nonconstructive question: is slide 8 of lecture 4 needed for the exam?

Resources on CANVAS

Slides will be regularly updated. If you notice typos, please e-mail me.
The discussion board is for asking constructive questions and answering them!

Nonconstructive question: is slide 8 of lecture 4 needed for the exam?
Answer: all content of the lectures & tutorials can be in the exam, and even with different numbers.

Resources on CANVAS

Slides will be regularly updated. If you notice typos, please e-mail me.

The discussion board is for asking constructive questions and answering them!

Nonconstructive question: is slide 8 of lecture 4 needed for the exam?

Answer: all content of the lectures & tutorials can be in the exam, and even with different numbers.

Positive participation will be noticed and can be later rewarded by references.

Assessment

Your mark for the module = a mid-term test (30%, multiple choice questions) in November + an end-of-term exam (70%) in January.

Assessment

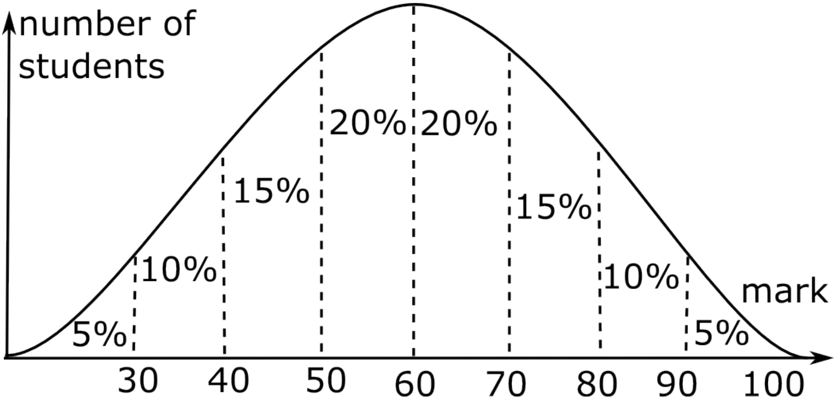
Your mark for the module = a mid-term test (30%, multiple choice questions) in November + an end-of-term exam (70%) in January.

Time: 2.5 hours, all questions will be marked.

Tutorial homework sets are formative (not contributing to the final mark), and will be a great preparation for your hand-written exam.

An example distribution of marks

The pass mark is 40, first class marks are 70+.



An expected average mark should be in [55,65].

Last years exam averages: ≈ 63, failure rate $\approx 10\%$.

Levels of understanding

Level 1 : a student thinks ‘I already know it’, e.g. all words are familiar. It’s a dangerous assumption.

Levels of understanding

Level 1 : a student thinks 'I already know it', e.g. all words are familiar. It's a dangerous assumption.

Level 2: a student is confused and asks a question. Your aim is to identify gaps in your knowledge.

Levels of understanding

Level 1 : a student thinks ‘I already know it’, e.g. all words are familiar. It’s a dangerous assumption.

Level 2: a student is confused and asks a question. Your aim is to identify gaps in your knowledge.

Level 3: a student has answered tutors’ questions at tutorials. Will you raise to this level in COMP229?

Levels of understanding

Level 1 : a student thinks ‘I already know it’, e.g. all words are familiar. It’s a dangerous assumption.

Level 2: a student is confused and asks a question. Your aim is to identify gaps in your knowledge.

Level 3: a student has answered tutors’ questions at tutorials. Will you raise to this level in COMP229?

Level 4: a student shares his knowledge with his peers. The real test for your understanding is to teach this concept to someone else.

Giving your feedback

For your questions and your feedback:

- e-mail O.Anosova@liverpool.ac.uk or
- post on the CANVAS discussion board

Giving your feedback

For your questions and your feedback:

- e-mail O.Anosova@liverpool.ac.uk or
- post on the CANVAS discussion board

The bottom line required by the university:
reply to student e-mails within 3 working days.

In the first 12 weeks I'll try to reply faster within few hours, the rest depends on your involvement.

Time to revise and ask questions

Revisions will be at the end of every lecture.

To benefit from the lecture, now you could

- prepare and ask your questions
- write down your summary in 2-3 phrases, e.g. list key concepts you have learned.

Time to revise and ask questions

Revisions will be at the end of every lecture.

To benefit from the lecture, now you could

- prepare and ask your questions
- write down your summary in 2-3 phrases, e.g. list key concepts you have learned.

Problem 1.2. Have you computed this value?

$51 \times 24 - 49 \times 93 + 27 \times 51 + 44 \times 49 = ?$

Final solution and summary

Solution 1.2. No calculator should be needed:

$51 \times 24 - 49 \times 93 + 27 \times 51 + 44 \times 49 =$

$51 \times (24 + 27) - 49 \times (93 - 44) = 51^2 - 49^2 =$

$(51 - 49) \times (51 + 49) = 2 \times 100 = 200.$

- COMP229 requires deep understanding.
- All the resources of COMP229 are on CANVAS.
- How to learn better: use all resources and share your knowledge with your classmates.