Imperial College
London

MATH96007 - MATH97019 - MATH97097

Methods for Data Science

2019-20

Prof Mauricio Barahona

The process of data science

- The science: what's a good question and what's the dataset you will use to answer it?
- What is the input and what are you trying to determine
- Data handling, cleaning and exploration: computer programming and knowing things
- THE MATH and the main analysis:
  - statistics, computing: what do you compute, why, and what does it mean?
  - machine learning: supervised, unsupervised..
  - networks: characteristics, analysis, comparisons
- Communication: words, visualisations, summaries to tell the story

Imperial College London

Aims and outcomes

- Learn mathematical concepts underpinning methods in learning from data
- The process of conceptualisation of the analysis
- The process of explanation of the analysis
- The mathematical justification of the analysis

- Getting a good exposure to current methods and their use in practice
- Challenges in dealing with (realistic) data

What the tools are, how they work mathematically, and how to analyse a dataset and clearly communicate the results

Imperial College
London

Data

Why do you need exploratory data analysis?

You will not have perfect data. EDA helps you to:

- find mistakes! (negative heights? copies of columns? unreasonable values? missing values?)
- check your assumptions
- get an idea of whehter the signal you want is actually in your data
  - if so, where?
  - maybe your problem is simple - just one predictor works
  - direction and size of relationships between predictors and outcome
- select your methods and tools
- find relationships between predictors

Data and types of EDA

Data: usually a big spreadsheet, table, data frame full of numbers and categories.

Typically one row per subject (patient, diamond, etc), and one column per variable (all the predictors and the outcome).

Reading 37 columns and 117,000 rows of a spreadsheet is not helpful.

We need to summarise and visualise the dataset in lots of ways to explore it, choose our method, find mistakes, etc.

Categorical data

- A *categorical* variable takes on one of a limited number of values (usually fixed)
- The distribution is really just the counts or frequencies of the different values
- Example: team name, which county someone lives in, breed of a dog
- Example: modules are mathematics, physics, computer science.
- Predicting a categorical variable is *classification*: we want to classify a new point into the right category.

Quantitative data

- A *quantitative variable* is numerical, and represents a measurable quantity
- Continuous examples: height, salary, stock price, profit, speed, blood count
- Ordinal examples (ordered): number of bedrooms, number of players, number of people in a city (nearly continuous for practical purposes)
- The distribution has a mean, variance, skewness etc.

Imperial College
London

Formalising the process

Data: Predictor variables

What might your *predictor* variables look like in practice?

- numbers
  - continuous measurements like height, price, profit, concentration
  - discrete measurements (integers): number of people, counts
- text: tweets, emails, patient records
- images: handwriting, medical images, photos
- even DNA sequences..
- combinations of these things

Randomness happens. It happens both in the *process* of getting the data, and in noise in the actual observations.

Data: outcome variables

What might your *outcome* (or "label") variables look like?

- Continuous numbers. In this case the problem is *regression*
- Categorical (success/failure, what type of fruit, what is in the image): in this case the problem is *classification*

Learning from data

Inputs belong to an input space: $x_i \in X$. For example, if you have p different continuous features (height, weight, grade on exam, peak exhaled air flow, blood hemoglobin...) for each data point, then $x_i \in R^p$.

x_i could also contain some other data that are not real numbers: county of residence, gender, company name...

Outputs y can be continuous or categorical: some number or feature or group, some knowledge about the data point.

Fundamentally, machine learning is about finding ways to link the predictors x_i to the outcomes y_i so that we can make new predictions.

Supervised learning:

- There is a known *outcome variable*: the truth.
  - emails where we we know which ones are spam
  - patients for whom we know their eventual outcome (heart attack or not)
  - patients where we know their glucose level
  - stocks and their prices 6 months later
  - tweets and their author (election team or Donald himself)
- We want to be able to predict that outcome for *new* observations of the input (predictor) variables

Unsupervised learning:

- There is not a known outcome variable that we want to predict
- Instead, we want to understand how the data are organised, clustered, related
  - use diverse measurements (blood, tumour shapes, gene expression) to identify similar types of breast cancers
  - explore differences in the bacterial communities in the guts of healthy individuals vs unhealthy
  - gain intuition for very high-dimensional data; make it more tractable

Ex: Are these questions for supervised or unsupervised learning?

1. I own a grocery chain and I use loyalty cards to track customer purchases. What are the major customer profiles? And what do they tend to buy together?
2. You have 6000 photos of dogs, cats, lizards and monkeys. Build a tool that can tell me which of these animals a newly-posted photo contains.

Supervised Learning

The goal is to create a function $y = f(x)$ relating *predictor* variables x to *outcome* variables y.

The data, called a *training set*, is a set of N input-output pairs, or data points:

$$x_i, y_i, \qquad i = 1, ..N$$

The goal is to be able to predict y_{new} from a new observation x_{new}.

Somehow we should be able to take uncertainty into account.

Classification vs Regression

Regression: predict a *quantitative* outcome (response) variable

- how much will increasing dosage affect the heart rate?
- how much does having an additional room increase the value of a house?
- does having a better shelf location impact sales of a product?

Classification: predict a *qualitative* outcome:

- Someone arrives at A&E with some symptoms. What medical condition do they have, of the ones that match the symptoms?
- Which emails are spam?
- Is a stock going to go up or down?

Imperial College
London

The basic process for supervised learning

1. Start with data (x_i, y_i), $i = 1, ..., N$. Each x_i usually a lot of features (components).
2. Decide: Classification or regression?
3. Define a *loss function* $L(y, f(x))$
4. Minimise the *mean sample loss*: $E(L) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$
5. Try to also have a reasonable *expected test loss*: $E(L(y_{new}, f(x_{new})))$

For example: in regression, the mean sample loss is the mean squared error:

$$L_{MSE}(y, f(x)) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

Imperial College
London

Now on the visualiser