Lecture 1: Introduction

MAST30034 Applied Data Science

Dr. Karim Seghouane
School of Mathematics & Statistics
The University of Melbourne

July 25, 2023

Outline

- Administrative Information
- Basics of Descriptive Statistics
- Introduction to Machine Learning
- Notions of Prediction
- Supervised and Unsupervised Learning
- Notions of Complexity

Administrative Information

Teaching staff

- Subject Coordinator and Lecturer: A/Prof. Karim Seghouane and Dr. Liam Hodgkinson.
- Instructor: Akira Wang.
- Tutors: Lucas Fern, Youran Zhou, Hong Yi Lin, and Dr. Ali Tebbi.

Administrative Information

Subject structure

- 1 hour lecture per week: Tuesday 11:00 to 12:00 in PAR-Elisabeth Murdoch G06
- 2 hour tutorial per week and 9 groups
  - Tutorials will be provided for the first 5 weeks but will only cover a portion of time. Weeks 6 to 12 will be for Project 2
  - Remaining time is intended for use on the assessments or to go through advanced learning content
    - Monday: G4. 09:00-11:00, G1. 13:00-15:00 and G3. 15:15-17:15 (PAR-Peter Hall-212, Nanson Laboratory),
    - Tuesday: G6. 15:15-17:15 (PAR-Peter Hall-G69, Thompson Lab),
    - Wednesday: G9. 11:00-13:00 (PAR-Peter Hall-G70 Wilson Laboratory),
    - Thursday: G2. 09:00-11:00 (PAR-Peter Hall-G70 Wilson Laboratory),
    - Friday: G7. 10:00-12:00, G8. 12:00-14:00 and G5. 14:15-16:15 (PAR-Peter Hall-212, Nanson Laboratory).

Administrative Information

Ongoing assessment

- 1 Individual Projects and two quizzes (30% + 10%)
- 1 Group Project (50%)
- 1 Team Review (10%)

Administrative Information

Subject pre-requisites and assumed knowledge

- Python and/or R
- Git (revision will be provided)
- LaTeX (revision guide will be provided)
- Any content covered in COMP30027, MAST30035, MAST20005, COMP20008

New tools

- Apache Spark 3.0 Framework (PySpark, dplyR/sparklyR)
- Apache Arrow Framework
- Geospatial Plotting libraries
- JupyterHub Server (bash, git)

Administrative Information

Project 1, quantitative analysis

- Already released, we recommend you start as soon as possible if you have yet to do so
- Worth 30% of your final grade due on the 13th of August at 11:59 am AEST (end of week 3)
- Requirements: Submit a GitHub repository and report written in LaTeX
- More covered in your tutorials this week
- LPT: Start as soon as possible

Administrative Information

Assignment 1, two set of quizzes

- Releases in Week 4 to 6
- Worth 10% of your final grade due soon after released
- Multiple choice questions
- The assignment will consist of questions related to the content taught in lectures 2 to 6

Administrative Information

Project 2, industry project

- Group Project starting Week 5 or 6 until Week 12
- Worth 50% of your final grade, though marks may be scaled within group members depending on the Team Reviews
- Groups must consist of 5 members and can be from any stream or tutorial delivery mode. However, you must find a suitable tutorial time to attend as Group Workshop Attendance is compulsory from Week 6 to 12
- There will be 2 to 3 projects to choose from; provided by our Industry Partners and/or from the University. There will be limited spots
- Groups may need to sign Non-Disclosure Agreements (NDAs). If you are unable to sign it, then you should choose the University provided project
- More information covered in Workshop 5/Lecture 6

Administrative Information

Project 2, team review

- A short Team Review and Self-Reflection due at the end of SWOTVAC (as there is no exam)
- Worth 10% of your final grade and will be your chance to rate yourself and your team members based on contribution and work effort
- Depending on the ratings, we may scale some group members up and down to ensure an equal amount of work has been done by all group members
- More covered closer to the date

Administrative Information

BYOD vs JupyterHub

- This year, we will be running a JupyterHub server supporting Python 3.8.3 and R 4.0.3 via Jupyter Notebooks or R-Studio
- It is strongly recommended students use their own devices if they have moderate specs (i.e 8-16GB of RAM, an i5 or equivalent, an external GPU if applicable)
- Access: https://mast30034.science.unimelb.edu.au/hub/login
- Log in using your UniMelb credentials (read more on the Canvas homepage)

Basics of Descriptive Statistics

Word cloud credit: Cal. State University

Statistics is a branch of mathematics dealing with the collection, organization, analysis, interpretation and presentation of data, which enable us to

- accurately describe the finding of scientific research,
- make decisions
- make estimation/prediction

Basics of Descriptive Statistics

Vector Data: is a vector of the same data type, each entry is an observation, the observations can be independent or dependent.
Rectangular Data: is the typical frame of reference for an analysis in data science is a rectangular data object, like a spreadsheet or database table. The key terms for rectangular data are

- Feature: A column in the table is commonly referred to as feature.
  Synonyms: attribute, input, predictor, variable.
- Outcome: many science projects involves predicting an outcome, often yes/no response or a summarized term.
  Synonyms: dependent variable, response, target, outcome.
- Record: A row in the table is commonly referred as a record.
  Synonyms: case, example, instance, observation, pattern, sample.

Basics of Descriptive Statistics

Some commonly seen data type:

- Continuous: "trip_distance", "trip_amount".
- Discrete: usually integer, "passenger_count" (Both continuous and discrete data are numerical.)
- Categorical: "blood_type"

Basics of Descriptive Statistics

Descriptive statistics: allow us to characterize the data based on its properties. There are four major types of descriptive statistics:

- Measures of frequency
- Measures of central tendency
- Measures of dispersion or variation
- Measures of association

Basics of Descriptive Statistics

Frequency: To be used when you want to show how often a response is given or show how often something occurs.

Central tendency: Mean, Median, and Mode

- Mean: the sum of a variables values divided by the total number of values.
- Median: the middle value of a variable.
- Mode: the value that occurs most often

Use this when you want to show how an average or the most commonly indicated response.

Basics of Descriptive Statistics

Dispersion or variation:

- Range = maximum - minimum.
- Variance or Standard Deviation: the difference between observed score and mean.

Use this when you want to show how spread out the data are. It is helpful to know when your data are so spread out that it affects the mean.

Basics of Descriptive Statistics

Measures of association between two variables: Covariance and Correlation

- A correlation coefficient is used to measure the strength of the relationship between numerical variables. The most common correlation coefficient is Pearson correlation coefficient, which can range from -1 to $+1$.
- Scatter plots can be useful (display the relationship between two quantitative or numeric variables by plotting one variable against the value of another variable).

Basics of Descriptive Statistics

Measures of association between multiple variables: Correlation matrix, Covariance matrix:

- A covariance (or correlation) matrix is a matrix with its j, k^{th} entry equals to the covariance (or correlation) between variable j and k.

$$\mathrm{Cov}(\mathit{trip\_dis}, \mathit{fare}, \mathit{tips}) \begin{bmatrix} 1 & 0.89 & 0.51 \\ 0.89 & 1 & 0.54 \\ 0.51 & 0.54 & 1 \end{bmatrix},$$

- Principle Component Analysis (PCA): PCA is a method to represent the original dependent (observed) variables using some new independent (latent) variables.

Basics of Descriptive Statistics

Model: Input $\rightarrow$ output Full model or reduced model? For a large

number of predictors (features/variables), we would like to identify the ones that exhibit the strongest effects to the response.

Basics of Descriptive Statistics

Some of the characteristics we might want to report of a dataset:

- Do the values tend to cluster around a particular point?
- Is there more than one cluster?
- How much variability is there in the values?
- How quickly do the probabilities drop of as we move away from the modes?
- Outliers: are there extreme values far from the modes?

Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand. Generalizing from our data to another set of cases is the business of inferential statistics.

Introduction to Machine Learning

- The wakeful human mind is constantly acquiring sensorial information from the environment, in the form of vision, hearing, smell, touch, and taste signals.
- The human mind is the best learning system there is to process these kind of data, in the sense that no computer as yet can consistently outperform a rested and motivated person in recognizing images, sounds, smells, and so on.
- Machine Learning applications in fields such as computer vision, robotics, speech recognition and natural language processing, generally have as their goal to emulate and approach human performance as closely as possible.

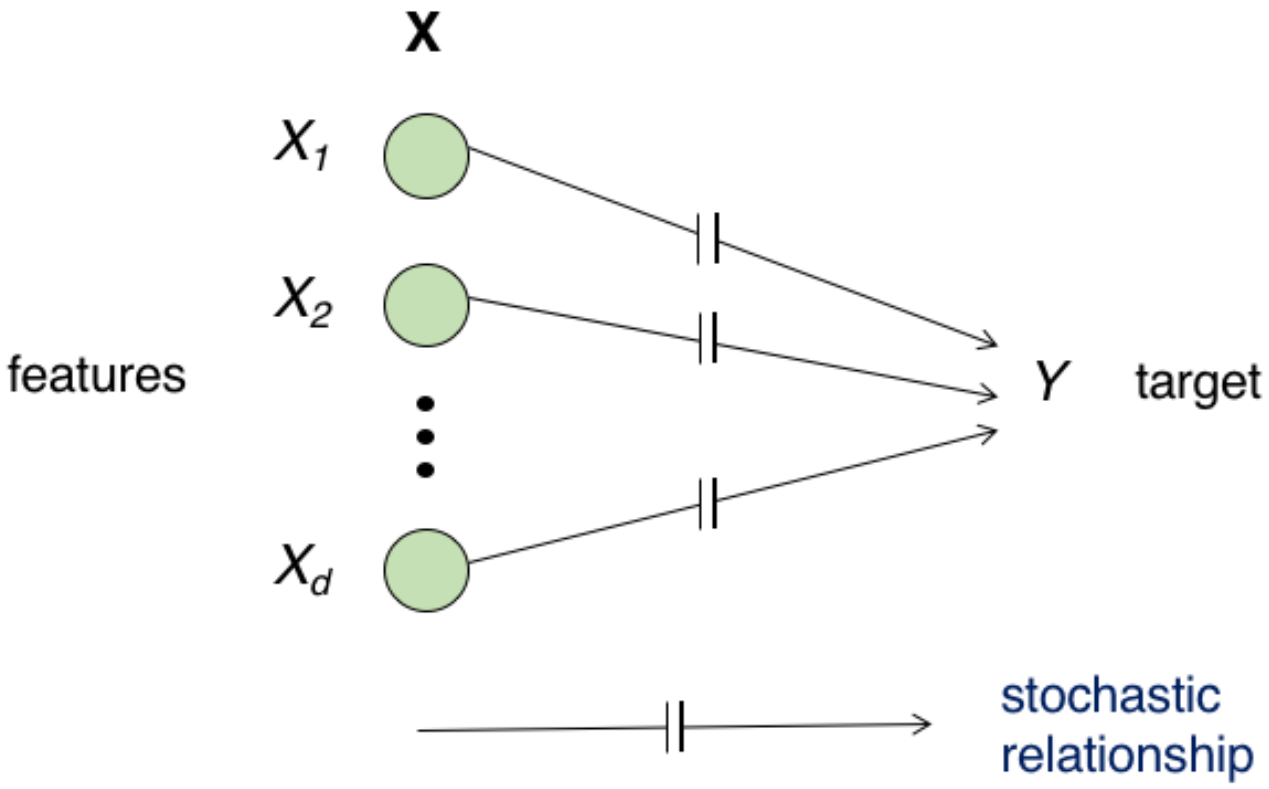
Introduction to Machine Learning

- Pattern recognition is one of the first engineering field in which AI problems were addressed. It dates back to 1960's.
- Machine Learning originated mainly in the neuroscience and computer science areas.
- Other identifiable areas closely related to this topic are Artificial Intelligence, Data Science and Signal and Image Processing.

Introduction to Machine Learning

- In supervised learning, information about the problem is summarized into a vector of measurements $\mathbf{x} \in \mathbb{R}^d$ also known as a feature vector, and a target $y \in \mathbb{R}$ to be predicted.
- The relationship between the feature vector $\mathbf{x}$ and the target y is, in practice, rarely deterministic, i.e., there is no function f such that $y = f(\mathbf{x})$.
- Instead, the relationship between $\mathbf{x}$ and y is better described by a joint probability distribution $p_{\mathbf{x},y}$.
- This state of uncertainty is due mainly to the presence of:
  - latent factors on which y depends but that are not observed or measured
  - measurement noise

Introduction to Machine Learning



Stochastic relationship between the features and target in supervised learning

Prediction

- A prediction rule or relationship produces a predictor $\psi : \mathbb{R}^d \to \mathbb{R}$, such that $\psi(\mathbf{x})$ predicts y. The predictor itself is not random.
- The construction of the predictor ψ uses information about the joint distribution $p_{\mathbf{x},y}$ which can be
  - direct knowledge about $p_{\mathbf{x},y}$
  - indirect knowledge about $p_{\mathbf{x},y}$ through a set of independent and identically distributed (i.i.d) sample $\mathbf{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, often called the training data
- The development of a predictor will employ a combination of these two sources of information.

Optimal vs. Data-Driven Predictors

- An optimal predictor $\psi^*(\mathbf{x})$ can be obtained in the case of complete knowledge of $p_{\mathbf{x},y}$ is available.
- Alternatively, a data-driven prediction rule must rely solely on $\mathbf{D}$.
- The optimal predictor can be obtained as $n \to \infty$ under certain conditions. The rate of convergence can however be slow.
- In the finite n case, some a priori knowledge about $p_{\mathbf{x},y}$ will help achieve good performance.

Predictor

- The performance and the development of a predictor is obtained using a loss function $\ell : \mathbb{R} \times \mathbb{R} \to R$. Examples include
  - the quadratique loss $\ell(\psi(\mathbf{x}), y) = (y - \psi(\mathbf{x}))^2$
  - the absolute difference loss $\ell(\psi(\mathbf{x}), y) = |y - \psi(\mathbf{x})|$
  - the missclassification loss

$$\ell(\psi(\mathbf{x}), y) = I_{y \neq \psi(\mathbf{x})} = \begin{cases} 1, & y \neq \psi(\mathbf{x}), \\ 0, & y = \psi(\mathbf{x}) \end{cases}$$

  where $I_\alpha = 1$ if α is true and $I_\alpha = 0$, otherwise.

- While we have access to a training data set, our interest is in the expected loss of ψ defined as

$$L[\psi] = E[\ell(\psi(\mathbf{x}), y)].$$

  The optimal predictor ψ^* minimizes $L[\psi]$ over all possibles $\psi \in P$, where P is the class of all predictors under consideration.

Supervised and Unsupervised Learning

- In supervised learning, the response y is available and defined. There are two types of supervised learning problems.
- **Classification** where $y \in \{0, 1, ..., K-1\}$ takes values in a finite alphabet, K is the number of classes. The variable y is in this case called a label to emphasize that it has no numerical meaning.
- in binary classification, there are two classes and $K = 2$.
- The used loss for classification is the missclassification loss with

$$\epsilon[\psi] = E\left[I_{y \neq \psi(\mathbf{x})}\right] = P\left(y \neq \psi(\mathbf{x})\right)$$

In binary classification $I_{y \neq \psi(\mathbf{x})} = |y - \psi(\mathbf{x})|$ or $(y - \psi(\mathbf{x}))^2$, they all yield the classification error $\epsilon[\psi]$.

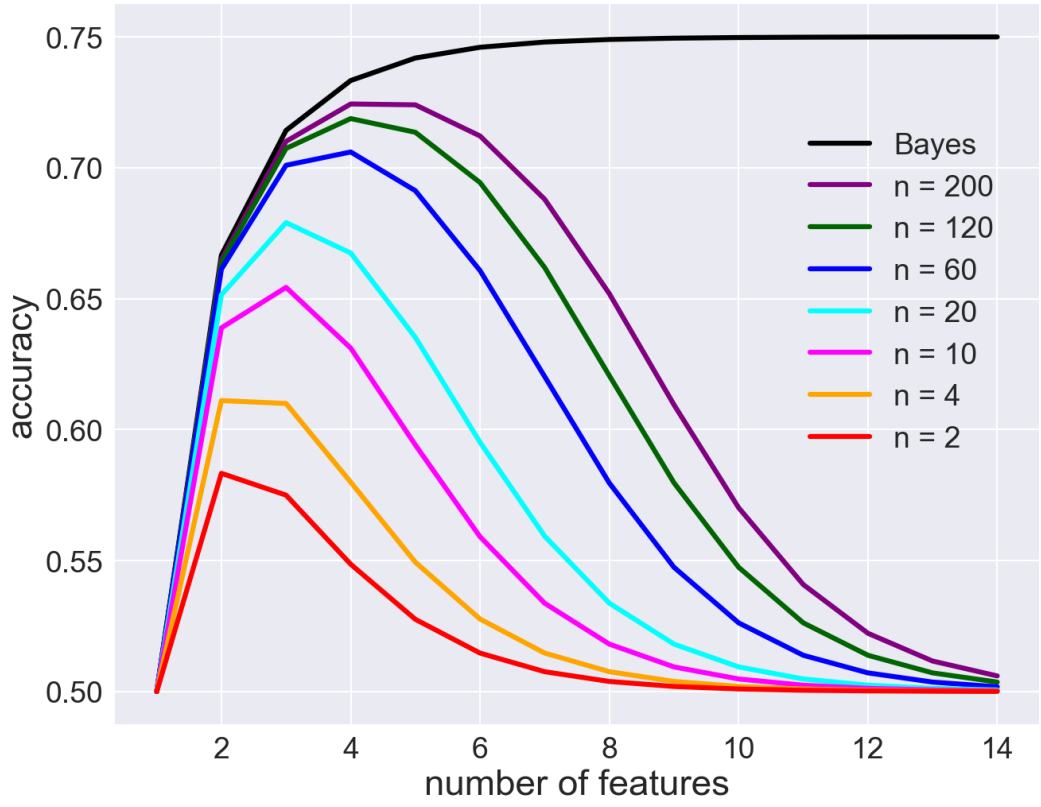
Supervised and Unsupervised Learning

- **Regression** where $y \in \mathbb{R}$, and the quadratic loss is a common used loss function and the prediction error $L[\psi]$ is called the mean-square error.
- Examples of regression include linear regression
- In **unsupervised learning**, y is not defined or given so only the distribution $p_{\mathbf{x}}$ is used.
- Examples of unsupervised learning operations include Principal Component Analysis (PCA).

Notions of Complexity

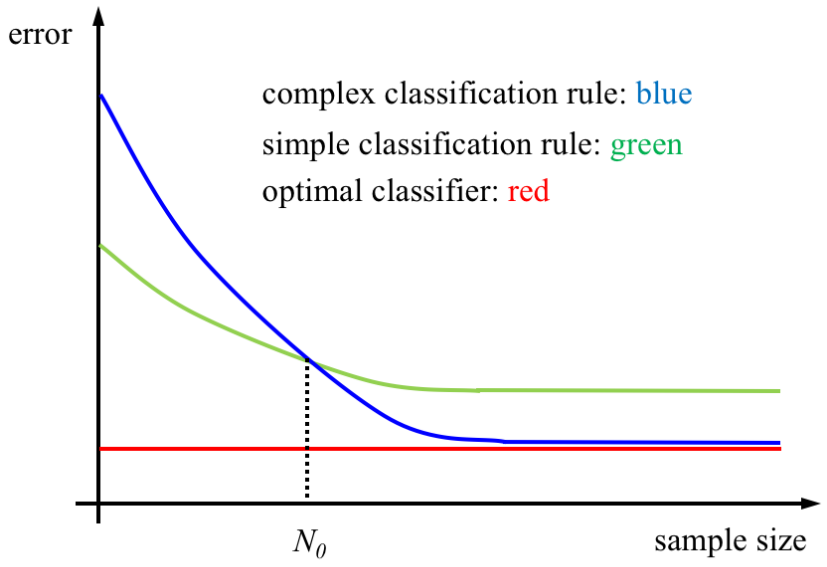
- Complexity trade-offs involving sample size, dimensionality and empirical performance. It is a characteristic feature of supervised learning methods.
- Notion of *curse of dimensionality*: for a fixed sample size, the expected classification error will improve by increasing the number of features, but eventually will decrease. This is a consequence of the large size of high-dimensional spaces, which require correspondingly large training sample sizes.
- *Scissors effect*: the expected error typically decreases as sample size increases, and more complex classification rules achieve smaller error for large sample sizes; however, simpler classification rules can perform better under small sample sizes, by virtue of needing less data.

Notions of Complexity



Expected accuracy in a discrete classification problem for various training sample sizes as a function of the number of predictors.

Notions of Complexity



Expected error as a function of sample size for two classification rules. There is a problem-dependent critical sample size N0, under which one should use the simpler classification rule.

Some References

- C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2010.
- G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning with Application in R*, Springer, 2013.
- U. Braga-Neto, *Fundamentals of Pattern Recognition and Machine Learning*, Springer, 2020.
- I. Koch, *Analysis of Multivariate and High-Dimensional Data*, Cambridge, 2014.
- M. P. Deisenroth, A. A. Faisal and C. S Ong, *Mathematics for Machine Learning*, Cambridge, 2020.