EMS702P Statistical Thinking and Applied Machine Learning

Week 1.1 – Introduction to Statistics and Machine Learning

Jun Chen

Introduction to Statistics and Machine Learning

Edition: v1.1

Table of Contents

1 Introduction to Statistics

Statistics is a collection of tools that you can use to get answers to important questions about data. You can use **descriptive statistical methods** to transform raw observations into information that you can understand and share. You can use **inferential statistical methods** to reason from small **samples** of data to whole domains (**population**).

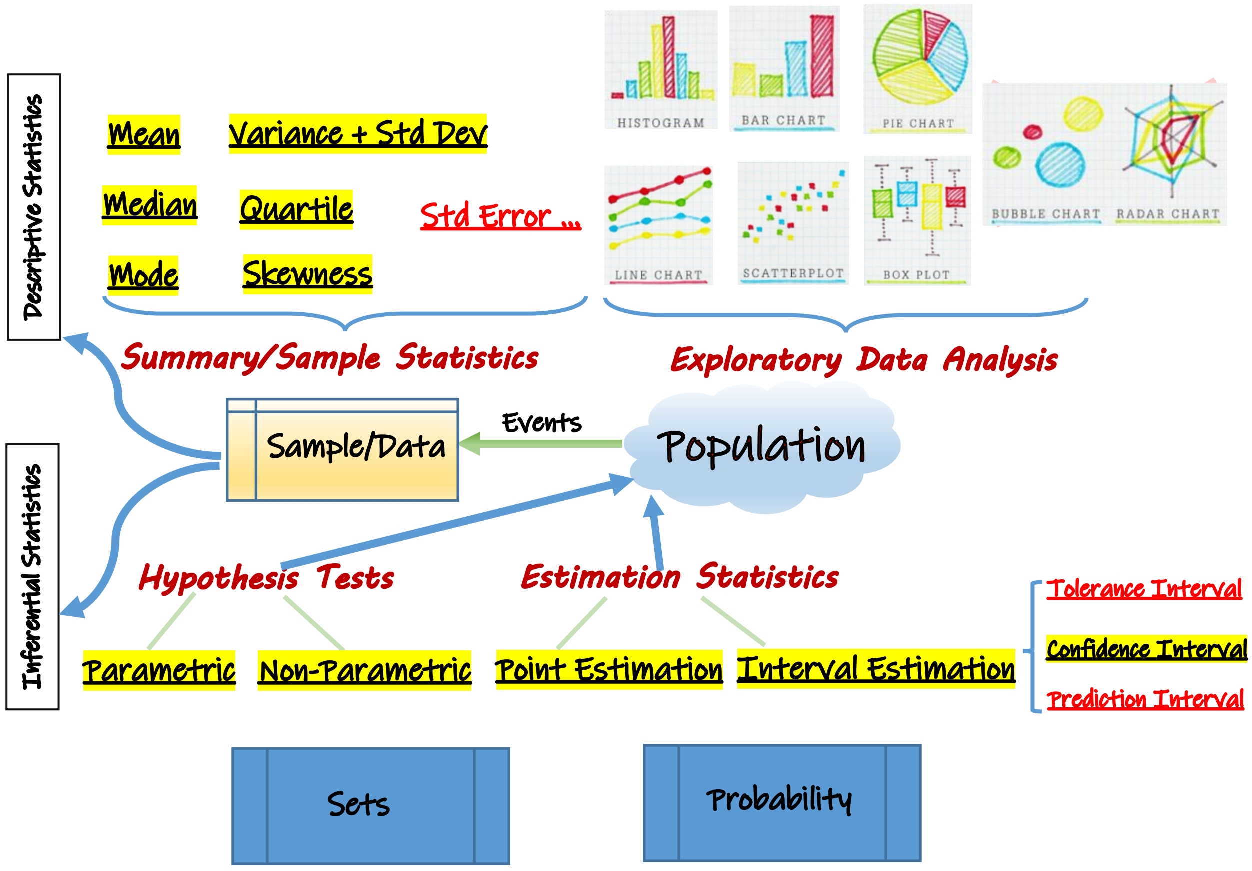
Statistical methods refer to a range of techniques from simple **summary statistics** intended to help better understand data, to **statistical hypothesis tests** and **estimation statistics** that can be used to interpret the results of experiments and predictions from models.

As a scientific method of data analysis, statistics has been applied throughout business, engineering and all of the social and physical sciences. Statistical methods are used at each step in an applied machine learning project. Machine learning practitioners have to **experiment**, analyse data and reach defensible conclusions about the outcomes of their experiments to determine how predictive models/algorithms behave when tested under real conditions.

Statistical methods will impact your practice of machine learning in the following ways:

- Use descriptive statistics and data visualizations to quickly and more deeply understand the shape and relationships in data and model skill (performance) and model predictions.
- Use inferential statistical tests to quickly and effectively quantify the relationships between samples, such as the results of experiments with different predictive algorithms or differing configurations.
- Use estimation statistics to quickly and effectively quantify the confidence in estimated model skill and model predictions.

1.1 The Map of Statistics



1.2 Terminologies

Experiment: An activity with an observable result, or set of results.

Outcome: An outcome is simply an observable result of an experiment.

Sample Space: A sample space is the set of all possible outcomes of an experiment.

Event: An outcome or set of outcomes to an experiment of interest to the experimenter.

Population: All members that belong to a certain group.

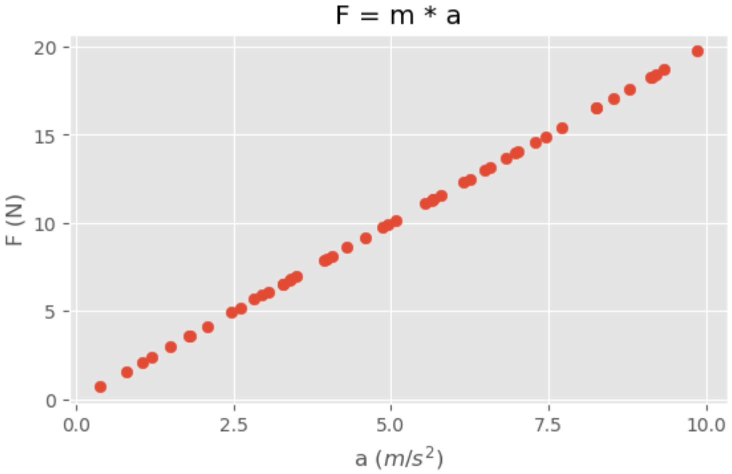
Sample: Some observed members of the population.

2 Introduction to Machine Learning

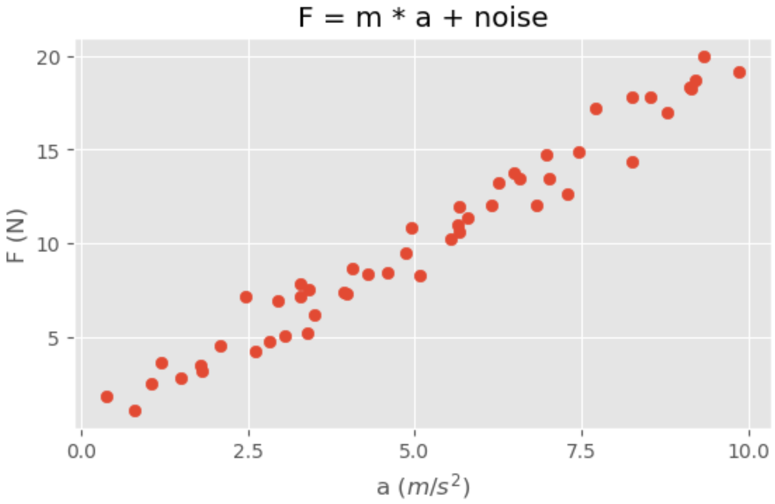
Machine learning is usually defined as the ability to acquire **knowledge**, by extracting patterns from raw **data**. It is a set of tools for modelling and understanding complex **datasets**. Here, data or datasets refer to (an) observed member(s), i.e. sample from a statistics perspective. **Noise** generally exists in data, causing data to differ from its expected or ideal behaviour (patterns). Uncovering (learning) patterns inside data is the general aim of machine learning.

Machine Learning: $Data = Patterns + Noise$

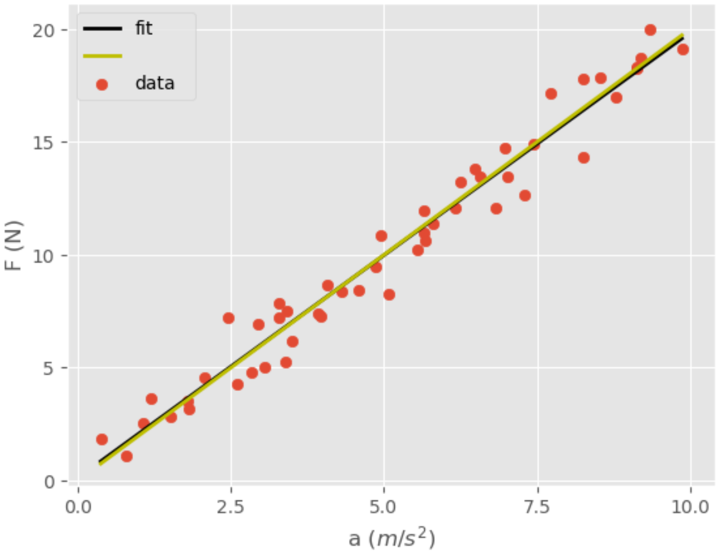
Using Newton's second law as an example, we plot the *ideal* Force vs. Acceleration of an object with $m = 2kg$, ignoring any imperfections due to, e.g. frictions, air drag, measurement errors, etc., we have



Adding those imperfections back to the force leads to data that a machine learning algorithm can learn from.



With linear regression (a machine/statistical learning algorithm you will learn in Week 6), we can find a very close approximation to $m * a$, even through the algorithm does not know it in beforehand.



In more general terms (or philosophical terms), machine learning echoes an old philosophical Problem of Universals: whether *Forms* (universals) exist in another world believed by Plato (left in the picture of the title page) or in this world believed by Aristotle (right in the picture of the title page). Here, universals are the things that two or more entities have in common and a universal has instances called particulars. Therefore,

Problem of Universals: $What\ we\ see = Universals + Particular\ properties$

Knowledge can take different forms such as

- Propositions (statement, law)

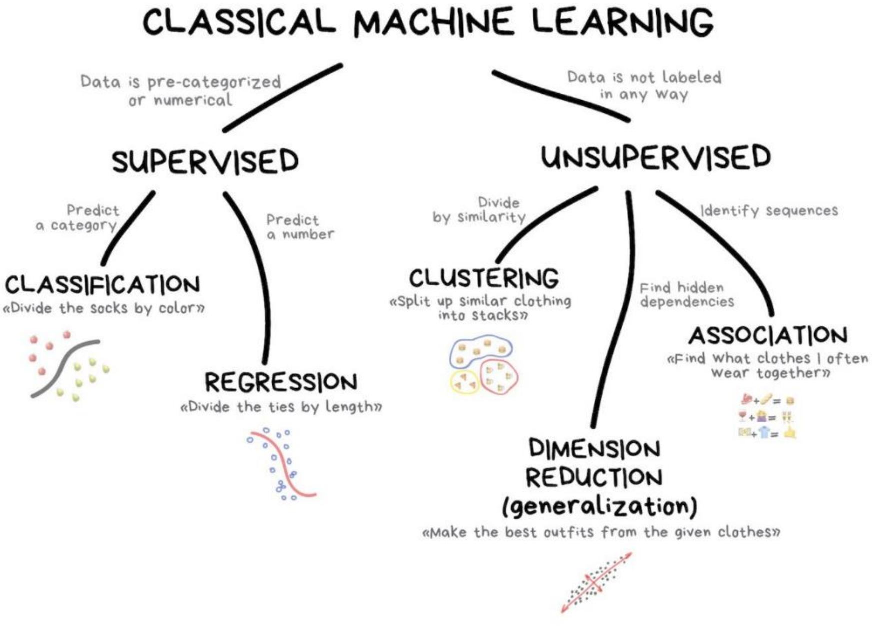
  Force is proportional to acceleration, and the ratio of the two is mass.
- Narratives (description, story)

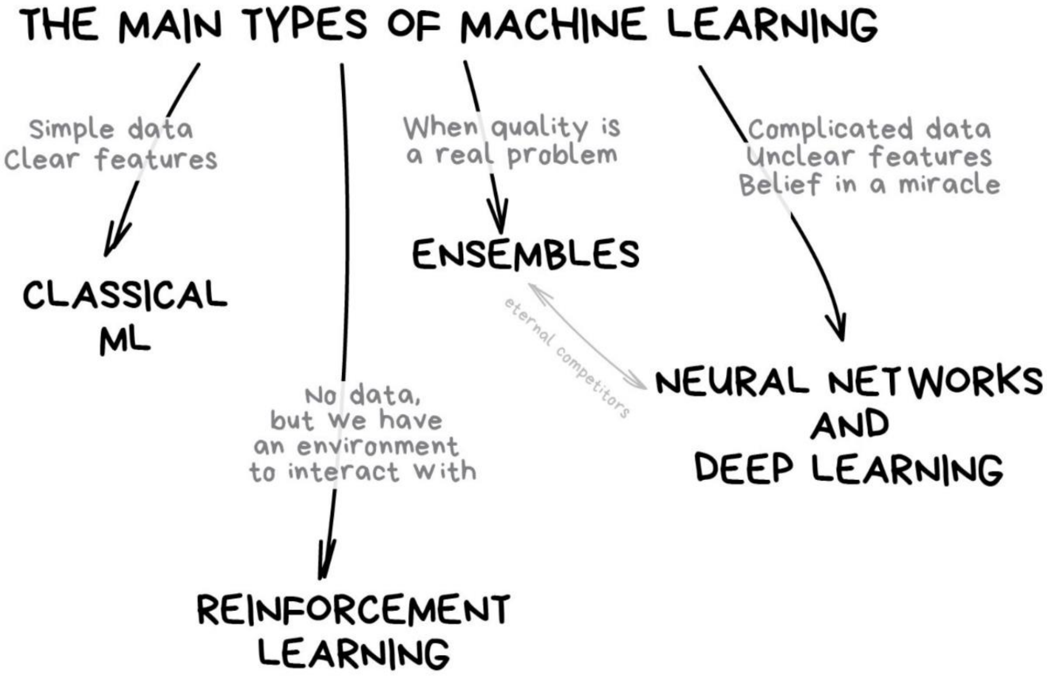
  With the same mass, in order to generate higher acceleration, we need more force.
- Models (Analytical, computational)

$$F = m * a$$

In this module, we focus on models, in particular computational models, as the representation of knowledge. Analytical models can be implemented, or sometimes approximated, by computational models. Computational models have mathematical expressions behind them although those mathematical expressions are not always equivalent to analytical models.

2.1 The Map of Machine Learning


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

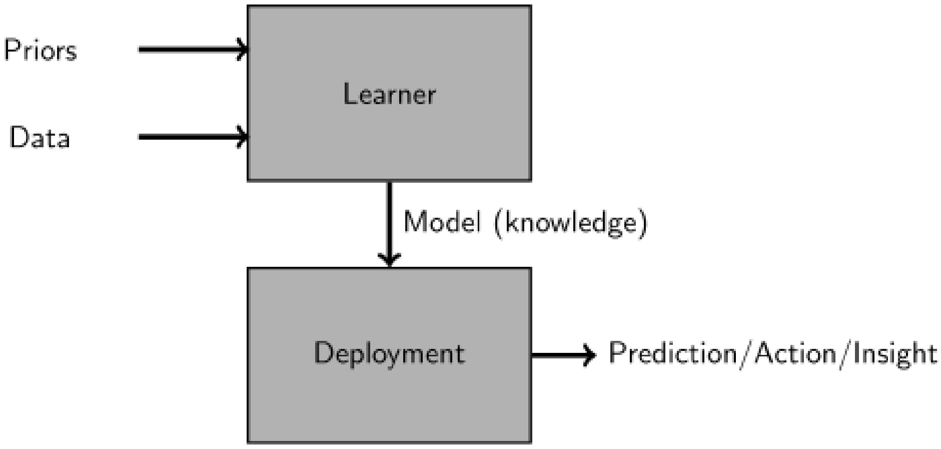


Source: https://vas3k.com/blog/machine_learning/

2.2 Machine Learning Basic Methodology

Models can be built, sold and deployed to deliver value. During the life of a model, we can distinguish two stages:

- **Learning stage:** The model is built.
- **Deployment stage:** The model is used.



In machine learning we are interested in finding the best model. Hence, we need a notion of model quality (skill, performance). However, our goal is to build models that work well during deployment, i.e. when presented with new data.

Basic machine learning methodologies include two separate tasks:

- **Training:** A model is created using data and a quality metric. We also say that we fit a model to a dataset.
- **Testing:** The performance of the model during deployment is assessed using new, unseen data.

Without rigorous methodologies, models are very likely to be of little use.

3 About EMS702P

3.1 Learning Goals

- Knowledge of Python to perform complicated mathematical calculations.
- Understanding of how eigenvalues and eigenvectors can be used in dimensionality reduction, and apply them to solve a wide range of engineering problems.
- Understanding of the concepts of probability and descriptive measures, the main features of a data set (exploratory data analysis), and design experiments.
- Construct confidence intervals for unknown parameters and test statistical hypotheses.
- Understanding of appropriate nonparametric methods and/or machine learning frameworks for a given engineering task using appropriate software tools.

- Evaluate the suitability of a machine learning algorithm to solve problems and formulate appropriate methodologies to evaluate the accuracy and robustness of machine learning models.
- Demonstrate an understanding of the broader challenge of engineering safety and effective human interaction with intelligent systems, such as gaining trust and explainable intelligence.

3.2 Module Contents and Schedules

W/k	Monday	Tuesday	Thursday		
	Lecture 11:00–13:00 Graduate Ctr: GC602	Lecture 16:00-17:00 Graduate Ctr: GC604	IT Class 9:00-10:00 Queens: LG5	PBL: 10:00-11:00 Graduate Ctr: GC203	Seminar 16:00-17:00 Queens: LG4
1	W1.1 Introduction	W1.2 Sets and Probability	Introduction to Python	In-class exercises for W1	
2	W2.1 Exploratory Data Analysis W2.2 Estimation Statistics I	W2.3 Estimation Statistics II	Python for W1.2 & W2.1	In-class exercises for W2	
3	W3.1 Hypothesis Test I W3.2 Hypothesis Test II	W3.3 PCA I	Python for W2.2 & W2.3	In-class exercises for W3	
4	W4.1 PCA II W4.1 Expert System, Fuzzy Sets & Operations	W4.2 Fuzzy Rules	Python for W3.1, W3.2, W3.3 & W4.1	CW1 Briefing	MATLAB in Machine Learning (MathWorks®)
5	W5.1 Fuzzy Inference – Mamdani W5.2 Fuzzy Inference – Sugeno	W5.3 Building a Fuzzy Expert Systems	Python for W5.1, W5.2 & W5.3	In-class exercises for W4 & W5	
6	W6.1 Linear Regression W6.2 Linear Classification I	W6.3 Linear Classification II	Python for W6.1, W6.2 & W6.3	In-class exercises for W6	Exploring ML in Life Critical Systems – Validation and Verification (NATS®)
7	Employability Week & Self Study				
8	W8.1 Neural Networks I W8.2 Neural Networks II	W8.3 Neural Networks III	Python for W8.1, W8.2 & W8.3	CW1 Presentation	
9	W9.1 Neural Networks Training I W9.2 Neural Networks Training II	W9.3 Time Series I	CW2 Briefing	In-class exercises for W8	
10	W10.1 Time Series II W10.2 Time Series Training	W10.3 Unsupervised Learning	Python for W9 & W10	In-class exercises for W9 & W10	
11	W11.1 Hybrid Intelligent Systems I W11.2 Hybrid Intelligent Systems II	W11.3 ML Practices	Python for W11.1 & W11.2	Drop-in Session for CW2	
12	W12.1 Deep Learning W12.2 Reinforcement learning	W12.3 Revision			

3.3 Structure of the Course

This module consists of **150** study hours (**96** hours of scheduled learning and teaching, including lectures, IT classes, problem solving classes, guided independent study and seminars + **54** hours of independent study, including assessment preparation, independent study time, etc). Its duration is **12** weeks (Week 7 is a Employability & Self Study week).

- **Lectures** (3h/week) will take place on-campus (see Schedules in Section 3.2).
- **Problem Solving Classes** (1h/week) will take place on-campus (see Schedules in Section 3.2).
- **IT Classes** (1h/week) will take place on-campus (see Schedules in Section 3.2). There are plenty of PCs in IT Labs. However, please feel free to bring your laptops with you. You may want to use this opportunity to set up the programming environment on your own laptop and programme more at home.
- **Seminars** (1h/session) consists of 2 guest lectures from leading industry experts and renowned external academics, which will take place in Weeks 4 and 6 on-campus (see Schedules in Section 3.2).

All learning materials (including the recorded sessions) are available on **QM+**, please use QM+ as the primary means for finding the course related information and communication. The forum on QM+ will be regularly monitored and is the best place for us to answer your questions, as it might be the question of other students.

Please make sure that the subjects of Email enquiries are formatted as follows "[EMS702P] [DESCRIPTIVE SUBJECT HERE>". Please make appointments for any 1-to-1 meetings, either on-Campus or through MS Teams.

3.4 Assessment

EMS702P is assessed in the following formats and sequence:

Type	% weight	sequence	Release Date	Submission
Coursework 1 (CW1)	15%	1	Thursday, Week 4	16 November 2023
Coursework 2 (CW2)	25%	2	Thursday, Week 4	14 December 2023
Exam	60%	3	TBC	-

A real-world application problem concerning artificial intelligence in Air Traffic Management(ATM) forms the Coursework that has two interconnected parts: CW1 and CW2. In CW1, you will create a dataset for CW2. In CW2, you will apply various machine learning algorithms covered in this module to build predictive models to predict aircraft taxi time that is crucial for efficient, safe and environmentally friendly ATM. In CW2, you will also select and apply statistical methods to evaluate model skills. More details of CW1 and CW2 can be found in the **EMS702P Coursework Student Pack**. It is strongly advised that you attend the briefing sessions for CW1 and CW2, as well as their drop-in sessions.

4 Further Readings

[1] Descriptive Statistics on Wikipedia.

https://en.wikipedia.org/wiki/Descriptive_statistics

[2] Statistical Inference on Wikipedia.

https://en.wikipedia.org/wiki/Statistical_inference

[3] Machine Learning for Everyone

https://vas3k.com/blog/machine_learning/