ST310 - Bits of maths - Week 1

September 2022

Introduction

The aim of these short notes will be to remind you of some important definitions, practices and results in probability and mathematics. Hopefully you will find that reviewing these concepts can help you in better understanding the material of this course.

Functions vs Models

In most first-year maths courses is introduced the definition of real function. A real *function* $f : \mathbb{R} \to \mathbb{R}$ is a way of assigning to each $x \in \mathbb{R}$ a specific and unique value $f(x) \in \mathbb{R}$.

The uniqueness bit of this definition is the most important: if we consider for example the function defined by $f(x) = x + 2$, we can uniquely determine the value $f(3)$ (which is $f(3) = 3 + 2 = 5$). It wouldn't be a function if we couldn't determine the value of f, or if we had more than one value to pick from.

Because we are used to write things like $y = f(x)$ in calculus, it might be confusing to see the notation: $Y = f(X) + \epsilon$ in a situation in which the same value of X can produce different outputs. Let us consider as an example the data set `House_prices` containing information about house prices in Budapest. Say that the only predictor variable of the data set is `sqm` (the square meters of the house), while the only response variable is `price`.

It could very well be that there are two distinct observations x_i, x_j such that the independent variable of the two observations is the same (that is, $x_i = x_j$) while the dependent variable of the two observations is different (that is, $y_i \neq y_j$). This is quite plausible, since all that it takes for this to happen is to have two houses with the same area but with different prices.

Therefore, we cannot say that `price` is a function of `sqm`. We can however use the language of random variables to write $Y \approx f(X) + \epsilon$ where $f(X)$ represents a function of the deterministic value X (in our case representing `sqm`), while ϵ represents a random variable (usually assumed to be a normal distribution with mean 0).

For a more detailed introduction to random variables, you can use one of the following resources:

https://www.probabilitycourse.com/chapter3/3_1_1_random_variables.php

Ch 1: https://math.dartmouth.edu/ prob/prob/prob.pdf

Ch III.2: https://ellerman.org/Davids-Stuff/Maths/Rota-Baclawski-Prob-Theory-79.pdf

Central Limit Theorem

Let us introduce the Central Limit Theorem with an example. Suppose we want to have a survey of the average height of the population of London, and so we open a stand in Trafalgar Square and start asking people for their height.

With the recovered data set, we can propose a guess for the average height of all the people in London. Indeed, we can create a list $X_1, X_2, \ldots, X_n$ of the heights of the people surveyed. And then we can calculate the average of the surveyed people: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Out intuition says that if we ask enough people, we will be able to produce a good guess about the average height of the London population. The Law of Large Numbers (LLN) tells us that our intuition is correct.

More precisely (*warning*, maths ahead), the LLN states that if we have an (infinite) sequence $X_1, X_2, \ldots, X_n, \ldots$ of independent and identically distributed real random variables with mean μ and variance σ^2, then we have

$$\lim_{n\to\infty} \bar{X}_n = \mu. \tag{LLN}$$

In our example we surely do not have an infinite amount of samplings of the population distribution, even if we might have that feeling. However, the idea that the height of two distinct people on the street is an independent random variable is reasonable, as it is the idea that both the observations come from the same distribution (the distribution of heights given by the people in London). Therefore, we have some mathematical argument to claim that our estimate will approximate μ (in the limit, no guarantees for small values of n).

If we want to know something more about our height distribution, the LLN is not going to be enough and we often use the Central Limit Theorem (CLT). Indeed, even in its weak form, the CLT is stronger than the LLN. In the same setting we presented above, CLT describes the behaviour of our approximation $\bar{X}_n$. Formally, this theorem states that the random variable $\sqrt{n}(\bar{X}_n - \mu)$ (which is a random variable because $\bar{X}_n$ is a function of the random variables $X_1, \ldots, X_n$) tends (for large numbers n) to the Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

This means that even if we blow-up the difference $\bar{X}_n - \mu$ by a factor $\sqrt{n}$ we still get a normal distribution with finite variance. Note that in the formula $\sqrt{n}(\bar{X}_n - \mu)$, only $\bar{X}_n$ is a random variable, while μ is a real number and n is the number of samples we took (that we assume is tending to infinity).

In summary, given an (infinite) sequence $X_1, X_2, \ldots, X_n, \ldots$ of independent and identically distributed real random variables with mean μ and variance σ^2, we have

$$\lim_{n\to\infty} \bar{X}_n = \mu, \tag{LLN}$$

$$\sqrt{n}(\bar{X}_n - \mu) \sim \mathcal{N}(0, \sigma^2). \tag{CLT}$$

For references: https://www.probabilitycourse.com/chapter7/7_1_0_limit_theorems.php
Ch 8, 9: https://math.dartmouth.edu/ prob/prob/prob.pdf
Ch IV.3: https://ellerman.org/Davids-Stuff/Maths/Rota-Baclawski-Prob-Theory-79.pdf