Lecture Notes Machine Learning

Felix Adam[a]

[a]*Barcelona Graduate School of Economics, Barcelona, Spain*

Abstract

This document contains my lecture notes for the course Machine Learning at the Barcelona Graduate School of Economics

1. Introduction

The main goal of this course is to develop an understanding for the reasons why certain machine learning algorithms work. Interestingly, some modern methods, such as Deep Learning, are not fully understood. It is not clear why these methods work or theory even says they shouldn't be as succesful.

We will start with discussing basic concentration inequalities, followed up by simple mean estimation. After that we'll start discussing supervised learning problems, mostly focused on classification with a minor detour towards regression. We then dive into the topic of empirical risk minimization and VC-theory. This will be followed by a discussion of linear classification, mostly support vector machines and kernel methods. Following, we transition to non-linear methods, especially classification trees and random forests. We finish the course with a discussion of clustering, spectral clustering and k-means and finally online-learning.

Email address: `felix.adam@bracelonagse.eu` (Felix Adam)

2. Mean Estimation

2.1. Motivation

We start the course with a seemingly simple task: estimating the mean of a population, given a sample drawn from the population.

The simplest considerable problem is to consider a setting where we are given independent, identically distributed (i.i.d) draws $X_1, X_2, ..., X_n$ of real-valued random variables. We further assume, that the mean (the expected value) exists $E[X] = m$. (Note that not all distributions have an expected value, such as the cauchy distribution).

Our goal is now, to find an estimate of m, based on the observed data.

An estimator is a function $m_n : \mathbb{R}^n \to \mathbb{R}$ that maps inputs to a value. We denote our estimate of the mean as the output of the function given the data $m_n(X_1, X_2, ..., X_n) = m_n$ (Note that the value of the estimate is commonly also denoted as m_n.

It is important to realize, that m_n is a function of random variables, so naturally, m_n is also a random variable. Ultimately, we would like to have an estimate (i.e., a data-based quantity) m_n that is close to the real mean m. We now need to figure out what "close" means.

2.2. Measuring the Error

A possible, and common way to measure the error is through the mean squared error (MSE)

$$\text{MSE} = E[(m_n - m)^2]$$

(Some terminology: The MSE is the risk of the estimator m_n under the squared loss.) However, the MSE is not the only possible measure of "closeness". Others are:

- Expected absolute error: $E[|m_n - m|]$
- Using probabilities: $P(|m_n - m| > \epsilon)$

We can also discuss the closeness in terms of loss functions $l : \mathbb{R} \to [0, \infty)$. The corresponding risk is the expected loss $E[l(m_n - m)]$. The loss functions associated with the discussed errors are:

- MSE: $l(x) = x^2$
- Absolute error: $l(x) = |x|$
- Probability : $l(x) = \mathbb{1}_{|x|>\epsilon}$

These criteria of closeness are not the same! So in order to assess an estimator, we first have to set a goal, in which sense do we want the estimator to be "good".

2.3. A simple Estimator

The most natural mean estimator is the **empirical mean**.

$$\overline{m_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

By the law of large numberers, as the sample size increases, the probability that the sample mean is equal to the true mean converges to 1.

Further, the sample mean is an unbiased estimate of the true mean, since:

$$E[m_n] = E[\frac{1}{n} \sum_{i=1}^{n} X_i] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = m$$