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:

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

(By the linearity of expectations).

We can now derive the MSE of the sample mean. Since we've shown that the estimator is unbiased, the MSE is the variance of the mean estimator.

$$E[(m_n - m)^2] = E[(\frac{1}{n}\sum_{i=1}^{n} X_i - m)^2] = \text{var}(m_n) = \text{var}(\frac{1}{n}\sum_{i=1}^{n} X_i)$$
$$= \frac{1}{n^2}\text{var}(\sum_{i=1}^{n} X_i) = \frac{\sigma^2}{n}$$

(By the linearity of the variance of independent random variables.) Note that this is only meaningful of the variance is σ^2 is fininte. Otherwise the MSE is also infinite.

What does this formula suggest? The error $|m_n - m|$ is typically of the order of $\frac{\sigma}{\sqrt{n}}$. This can be derived from the following observation:

We know from the properties of the variance (not smaller than 0) that:

$$(E[X])^2 \leq E[X^2]$$

since $\text{Var}(X) = E[X^2] - (E[X])^2$. From this follows, that $E[X] \leq \sqrt{E[X^2]}$ and thus

$$E[|m_n - m|] \leq \sqrt{E[m_n - m)^2]} = \frac{\sigma}{\sqrt{n}}$$

Thus, we've established a first bound for the MSE of the sample mean. We can say that the expected distance between the sample mean and the true mean depends on the variance and the sample size.

Very often one needs control probabilities of the type

$$P(|m_n - m| > \epsilon)$$

This will especially be important when wee need to estimate the mean of not just one but many random variables say N of them and we want to make sure that the errors are simultaneously small. In other words, we need to control

$$max|m_n^{(j)} - m^{(j)}|$$

In order to deal with these types of probabilities and find proper bounds, we'll need to establish some common inequalities.

2.4. Inequalities for sums of indenpendent random variables

2.4.1. Markov's Inequality

We start with the Markov inequality. The inequality yields an upper bound for the probability that a random variable X exceeds a given value t. We are using the special case of the Markov inequality where $X \geq 0$. Then

$$P(X \geq t) \leq \frac{E[x]}{t}$$

Proof

Let $\mathbb{1}_t$ be the indicator function that event t occurs. We have $\mathbb{1}_{(X \geq t)} = 1$ if $X \geq t$ and $\mathbb{1}_{(X \leq t)} = 0$ otherwise. Then given that $t > 0$ we find

$$t\mathbb{1}_{(X \geq t)} = X$$

since if $X < t$ then $\mathbb{1}_{(X \leq t)} = 0$ and so $t\mathbb{1}_{(X > t)} = 0 \leq X$. Othweise, if$X \geq t$, we have $\mathbb{1}_{(X \leq t)} = 1$ and thus $t\mathbb{1}_{(X > t)} = t \leq X$. Taking the expecations on both sides:

$$E[t\mathbb{1}_{(X \geq t)} \leq E[X]$$

using

$$tE[\mathbb{1}_{(X \geq t)}] = a(1 \cdot \mathrm{P}(X \geq t) + 0 \cdot \mathrm{P}(X < t)) = t\mathrm{P}(X \geq t)$$

Thus we have

$$tP(X \geq t) \leq E[X]$$

2.4.2. Chebyshev's Inequality

Chebyshev's inequality bounds the probability that a random variable deviates from it's expected value by more than a given threshold by the variance of the random variable itself.

$$P(|X - E[X]| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$$

Proof

The proof can be derived by using Markov's inequality and by transforming the inequality through taking the square.

$$P(|X - E[X]| \geq t) = P((X - E[X])^2 \geq t^2) \leq \frac{(E[X - E[X])^2}{t^2} = \frac{\mathrm{Var}(X)}{t^2}$$

In particular, for the sample mean we find:

$$P(|m_n - m| \geq \epsilon) \leq \frac{\mathrm{Var}(m_n)}{t^2} = \frac{\sigma^2}{n\epsilon^2}$$

This implies the weak law of large numbers. The probability that the difference between the sample mean and the true mean is larger than a given ϵ converges to zero as the sample size grows, for all $\epsilon > 0$.

2.4.3. Chernoff Bounds

The Chernoff bound describes exponentially decreasing bounds on tail distributions of sums of indenpendent random variables. It is a sharper bound than Markov's ineqaulity and Chebyshev's inequlaity.

We use the Chernoff bound since we would like to have sharper bounds for $P(X - E[X] \geq t)$.

The Chernoff bound for a random variable X is attained by applying the exponential function:

$$P(X - E[X] \geq t) = P(e^{\lambda(X-E[X])} \geq e^{\lambda t)}) \leq \frac{E[\exp(\lambda(X - E[X])]}{\exp(\lambda t)}$$

For any $\lambda > 0$. The function $E[e^{tX}]$ of the random variable X is called the **moment generating function**. So the probability that X exceeds it's expected value by t is bound by the moment generating function. We can now bound the moment generating function and optimise the bound in *lambda*.

(Example)

Taking a random variable $X \sim N(0, 1)$ we want to derive $P(X \geq t)$. The moment generating function is $e^{\lambda^2/2}$. We therefore have

$$P(X \geq t) \leq \frac{e^{\lambda^2/2}}{e^{\lambda t}}$$