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. Concentration Inequalities

Concentration inequalities provide bounds on how a random variable deviates from some value (here often it's expected value). These inequalities can be sorted according to how much information about the random variable is needed in order to use them.

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. Using Markov's inequality we don't need any information about the random variable, expect that it's expected value exists.

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

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

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

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

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. Thus, the use of Chebyshev's inequality requires information about the variance.

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

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 λ.

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

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

Optimizing this w.r.t λ we get that $\lambda = t$.

2.5. Bounds for the Mean Estimator

Having discussed these bounds, we can now apply them to find bounds for the mean estimator discussed before. We want to get an upper bound for the probability that our mean estimator deviates from the true mean by more than a value t.

Let $X_1, ..., X_n$ be i.i.d random variables with mean m. Using Chernoff bounds we find:

$$P(m_n - m \geq t) = P(\sum_{i=1}^{n} X_i - mn \geq nt) \leq \frac{E[exp(\lambda \sum_{i=1}^{n} X_i - m)]}{exp(\lambda nt)}$$

In the numerator, we drop nm since we can pull the m into the sum n times. To further simplify this upper bound for the mean we can simplify the numerator:

$$\begin{aligned} E[e^{\lambda \sum_{i=1}^{n}(X_i - m)}] &= E[\prod_{i=1}^{n} e^{\lambda(X_i - m)}] \\ &= \prod_{i=1}^{n} E[e^{\lambda(X_i - m)}] = (E[e^{\lambda(X_1 - m)}])^n \end{aligned}$$

The first equality is due to the properties of exponents, the second transformation is due to the independence of the X_i and the third is due to the fact that the X_i are identical.

Using this result we find:

$$P(m_n - m \geq t) \leq \frac{(E[e^{\lambda(X_1 - m)}])^n}{e^{\lambda nt}}$$

The probability that the estimated mean deviates from the true mean by t is bound by the moment generating function and the number of observations n.

2.6. Hoeffding's Lemma

Hoeffding's lemma is an inequality that bounds the moment generating function of any bounded random variable.

Let X be any real valued random variable that is bounded in the interval $[0, 1]$. Then

$$E[e^{\lambda(X-E[X])}] \leq e^{\frac{\lambda^2}{8}}$$

Hoeffding's lemma shows, that any bounded random variable is a subgaussian. In general for any interval $[a, b]$ we get:

$$E[e^{\lambda(X-E[X])}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}$$

2.7. Hoeffding's Inequality

Hoeffding's inequality provides an upper bound on the probability that the sum of bounded independent random variables deviates from its expected value by more than a ceratin amount.

Let $X_1, ..., X_n$ be indepdent, random variables taking values in $[0, 1]$ with $E[X] = m$. Then the moment generating function of the sum of these variables is bounded by Hoeffding's lemma.

$$E[e^{\lambda \sum_{(i=1)}^{n}(X_i-m)}] = \prod_{(i=1)}^{n} E[e^{\lambda(X_i-m)}] \leq e^{\frac{n\lambda^2}{8}}$$

Going back to the problem of estimating the mean from data we get (by applying Chernoff bounds):

$$P(m_n - m \geq \epsilon) \leq \frac{e^{n\lambda^2/8}}{e^{\lambda n\epsilon}} = \exp(n[\lambda^2/8 - \lambda\epsilon])$$

We can minimize this bound with respect to λ. The optimization yields $\lambda = 4\epsilon$. Plugging this in to the formula we get **Hoeffding's inequality**.

$$P(m_n - m \geq \epsilon) \leq e^{-2n\epsilon^2}$$

For this bound to hold, we don't need much, just independence between the random variables and that they are bounded by $[0, 1]$. Hoeffding's inequality therefore gives non-asymptotic and distribution free bound for the distance of the mean estimate from the true mean.

2.8. Bernstein's Inequality

Hoeffding's inequality is elegant and easy to use but (because of it's distribution-free nature) it is necessarily not tight for some distributions. In particular, the dependen on the variance is missing from the exponent. Using the Chernoff bound, one may prove such a bound, called Bernstein's inequality.

Let $X_1, ..., X_n$ be independent random variables such that $X_i \leq 1$, $E[X_i] = 0$ and $\mathrm{Var}(X_i) = \sigma^2$ (bounded by 1, zero mean and finite variance). Then for some $b > 0$

$$P(\sum_{i=1}^{n} X_i \geq t) \leq \exp[\frac{-t^2}{2(\sigma + \frac{bt}{3})}]$$

By symmetry we can also say that

$$P(m_n - m \leq -\epsilon) = P(m - m_n \geq \epsilon)$$

Accordingly, the probability that the absolute error is bigger than ϵ is

$$P(|m_n - m| \geq \epsilon) \leq 2e^{-2t^2/n}$$

2.9. The Union Bound

The union bound, also known as Boole's inequality, says that for any finite or countable set of events, the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events.

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}\left(A_i\right)$$

2.10. The Median of Means Estimator

Having discussed the various bounds we can now assess the upper bound for the empirical mean.

Given i.i.d data $X_1, ..., X_n$ with expected value $E[X]] = m$ and finite variance, then using Chebyshev's inequality we find

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

this is the best assessment we can make, since we can't assume that the data is bounded. The question is now, whether there is a better estimator than the empirical mean. We will now introduce such an estimator, called the median of means estimator.

2.10.1. Median of Means

We divide the data $X_1, X_2, ..., X_n$ into k blocks. For simplicity we assume that $n = km$. So block one will be $X_1, ..., X_m$, block two will be $X_{m+1}, ..., X_{2m}$ and so on up to block k.The median of the means estimator computes the empirical mean μ_i in each block. We then compute the median of these means $(\hat{m_n})$ to obtain the estimate of the mean of the full data. We can now analyse the quality of this estimation. For each single estimate μ_i we know that by Chebyshev's inequality:

$$P(|\mu_i - m| \geq \frac{2\sigma}{\sqrt{m}}) \leq \frac{1}{4}$$

Proof:

Using Chebyshev's inequality we can write:

$$P(|\mu_i - m| \geq \frac{2\sigma}{\sqrt{m}}) \leq \frac{\mathrm{Var}(\mu_i)}{\frac{2\sigma}{\sqrt{m}}}$$

The variance of the mean estimator can be derived using the linearity property of the variance of i.i.d data.

$$\begin{aligned}\mathrm{Var}(\mu_i) &= \mathrm{Var}(\frac{1}{m}\sum_{i=1}^{m} X_i) \\ &= \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}(X_i) \\ &= \frac{\sigma^2}{m}\end{aligned}$$

Using the variance, we arrive at the equation shown above.

Now the question is, how good is the median of the means estimate? By the property of the median, if $|\hat{m_n} - m| > \frac{2\sigma}{\sqrt{m}}$ then it must be that at most $k/2$ of the $\hat{\mu_i}$ are such that $|\hat{\mu_i} - m| > \frac{2\sigma}{\sqrt{m}}$. The probability of this is at most

$$P(\mathrm{B}(K, 1/4) \geq k/2) = P(B - k/4 \geq k/4) \leq e^{-2k/16}$$

by Hoeffding's inequality (where B is a binomial random variable). Fixing the allowed eror δ to be $\delta \in (0, 1)$ then $e^{-k/8} \leq \delta$ implies $k \geq 8\log(1/\delta)$. So we choose the number of blocks to be $k = \lceil 8\log(1/\delta)\rceil$ depending on our allowed erro. Then $m = \frac{n}{k} = \frac{n}{\lceil 8\log(1/\delta)\rceil}$ and we obtain the following.

Theorem:

Fix $\delta \in (0, 1)$. The mediean of means estimator with $k\lceil 8\log(1/\delta)\rceil$ blocks satisifies

$$P(|\hat{m_n} - m| \geq \frac{2\sigma\sqrt{8\log(1/\delta)}}{\sqrt{n}}) \leq \delta$$

The median of means estimator has sub-Gaussian performance under the only condition that the variance is finite. This is much better than the empirical mean. The estimate depends on δ, for each δ we have a different estimate.

3. Random Projections for Dimensionality Reduction

We can illustrate the power of concentration inequalities (in particular, Chernoff bounds) by a suprising dimensionality reduction technique that has many applications.

Let $a_1, ..., a_n \in \mathbb{R}^D$ where D is large (in genetics for example). We would like to find a mapping $f : \mathbb{R}^2 \to \mathbb{R}^d$ such that f preserves pairwise distances in the sense that for all $i, j = 1, .., n$

$$||f(a_i) - f(a_j)|| \approx ||a_i - a_j||$$

In other words, we'd like to represent these points in a lower dimension space, without loss of information. If exact equality is required, there's not much one can do (unless all points fall in a low-dimensional subspace of $\mathbb{R}^D$). However, if we allow some error, the situation changes dramatically.

Let $\epsilon > 0$. We require that for all $i, j = 1, ..., n$

$$1 - \epsilon \leq \frac{||f(a_i) - f(a_j)||^2}{||a_i - a_j||^2} \leq 1 + \epsilon$$

In other words, we want the ratio of euclidian distances to be close to one, with error ϵ. Surprisingly, such an $f : \mathbb{R}^2 \to \mathbb{R}^d$ exists, whenever $d \geq \frac{8 \log(n)}{\epsilon^2}$ independently of how large D is. This is also known as the Johnson-Lindenstrauss-Lemma and used in compressed sensing, dimensionality reduction and graph embedding.

It is perhaps suprising how easy it is to find such functions. In fact, we can take f to be linear, that is, f is of the form $f(a) = Wa$ where $W = (W_{ij})_{dxD}$ is a matrix. How do we find such a matrix? We can pick it at random! We show that if $W_{ij} \sim N(0, 1/d)$ all of them independent, then the matrix W has the desired property, with high probability.

We note: $f(a_i) - f(a_j) = W(a_i - a_j) = Wb_{ik}$. For any vector $b \in \mathbb{R}^D$:

$$
\begin{aligned}
E[||Wb||^2] &= E[(\sum_{i=1}^{d}\sum_{j=1}^{D} b_j W_i j)^2] \\
&= \sum_{i=1}^{d} E[(\sum_{j=1}^{D} b_j W_{ij})^2] \\
&= \sum_{i=1}^{d} E[\sum_{j=1}^{D} \frac{1}{d} b_j^2] \\
&= \frac{d}{d}\sum_{j=1}^{D} b_j^2 = ||d||^2
\end{aligned}
$$

because the W_{ij} are independent and have variance $\frac{1}{d}$.

Short info for the third step in the derivation: We know that $\sum b_i N_i \sim N(0, \frac{\sum b_i^2}{d})$ if $N_i \sim N(0, 1/d)(i.i.d)$. We now want to find $E[(\sum_{j=1}^{D} b_j W_{ij})^2]$. We can make use of the variance:

$$
\mathrm{Var}(X) = E(X^2) - (E(X))^2
$$

where $X \sum_{j=1}^{D} b_j W_{ij}$. In this case we know that $(E(X))^2 = 0$, by independence. Therefore, $E(X^2) = \mathrm{Var}(X) =, \frac{\sum b_i^2}{d}$

In particular, for any a_i, a_j

$$
E[||f(a_i) - f(a_j)||^2] = E[||(a_i - a_j)||^2] = ||a_i - a_j||^2
$$

So one needs to show that, with high probability

$$
\max_{i,j=1,\ldots,n} \left| \frac{||W(a_i - a_j)||^2}{||a_i - a_j||^2} - 1 \right| < \epsilon
$$

We can re-write:

$$\frac{||W(a_i - a_j)||^2}{||a_i - a_j||^2} = \left|\left|W\frac{a_i - a_j}{||a_i - a_j||}\right|\right|^2 = ||Wb_{ij}||^2$$

Where b_{ij} is a unit vector. But for any unit vector $b \in \mathbb{R}^D$,

$$||Wb||^2 - 1 = \sum_{i=1}^{d}(\sum_{j=1}^{D} W_{ij}b_j)^2 - 1 = \sum_{i=1}^{d}(N_i^2 - E(N_i^2))$$

Where $\sum_{j=1}^{D} W_{ij}c_j = N_i$.

This can be shown by using the fact that are i.i.d $W_{ij} \sim N(0, \frac{1}{d})$. Due to independence $E[\sum_{j=1}^{D} W_{ij}b_j] = \sum_{j=1}^{D} b_j E[W_{ij}] = 0$. The variance is $\text{Var}(\sum_{j=1}^{D} W_{ij}b_j) = \sum_{j=1}^{D} b_j^2 \text{Var}(W_{ij}) = \text{Var}(W_{ij}) = \frac{1}{d}$ since b is a unit vector and for any unit vector b, $\sum_{j=1}^{D} b_j^2 = 1$. Furthermore, $\sum_{i=1}^{d} E[N_i^2] = \sum_{i=1}^{d} \frac{1}{d} = 1$.

Note, that $\sum_{i=1}^{d} N_i^2 \sim \mathcal{X}^2(d)$ is a sum of squared random normal variables. We may now prove, using the Chernoff bound, that

$$P(|||Wb||^2 - 1| > \epsilon) \leq e^{-d\epsilon^3/4}$$

and therefore, by the union bound, since there are $\frac{n(n-1)}{2} < n^2$ points a_i, a_j

$$P\left(\max_{i,j=1,\ldots,n}\left|\frac{||W(a_i - a_j)||^2}{||a_i - a_j||^2}\right| > \epsilon)\right) \leq n^2 e^{-\epsilon^3 d/4}$$

Whenever $d > \frac{4\log(n^2/\delta)}{\epsilon^2}$.

4. Binary Classification

4.1. The Classification Problem

We are all familiar with the issue of binary classification, say classifying animals into cats and dogs, depending on their weight and tail length. However, in order to assess whether we are doing the right thing when we apply a model, we need a mathematical framework. This section will help to develop an understanding of such a framework.

Usually, the starting point of (supervised) binary classification is a set of observations, in our case denoted by χ. We want to assign each of these observations $x \in \chi$, to one of the two classes, say 0 or 1. Formally, a classifier is a function $g : \chi \to \{0, 1\}$. The problem can be set up as a statistical hypothesis testing problem. In this framework, one assumes that the observation X is random, and it has different (conditional) distributions depending on which of the two classes it belongs to. Formally, (X, Y) is a pair of random variables taking values in $\chi x\{0, 1\}$. X represents the observation and Y its class or label.

To make the further discussion more concise, we can establish some properties of the point distribution of (X, Y). Most importantly, we want to make statments about the distribution of our parameters or features X which we will call μ and the a posteriori probabilities of Y given X, which we'll call η. Additionally, we can describe the a priori probabilities of an observation belonging to either class i by q_i and the conditional distribution of X depending on Y as the class conditional distributions. In general we may say that, X can take values in a subset of χ which we'll call A.

- $\mu(A) = P(X \in A)$ for all $A \subset \chi$
- $\eta : \chi \to [0, 1]$, $\eta(x) = P(Y = 1|X = x)$.
- $q_1 = P(Y = 1), q_0 = P(Y = 0)$
- $P(X \in A|Y = 0), P(X \in A|Y = 1)$ for $A \subset \chi$

4.2. Bayes Risk and Classifier

Having set up the problem formally, we now want to make statments about the quality of a classifier or a prediction rule. In order to do so, we first need to find a baseline classifier, to which we can compare other classifiers. Ideally, this classifier has the optimal decision rule! So what is the optimal decision rule? Let's say for now, that we want to minimize the probability of error. Formally we want to **minimize the risk** $R(g)$ of classifier g, given by its probability of error.

$$R(g) = P(g(x) \neq y)$$

Let's further assume, that we measure errors in a symmetric way.

$$R(g) = P(g = 1|Y = 0)P(Y = 0) + P(g = 0|Y = 1)P(Y = 1)$$

This may not be natural (or wise) in some applications. For more general cases, we may have a loss function l, which measures the different types of errors.

$$l : Y \times Y \to R$$

Given a loss function, the risk is the expected value of said function $R(g) = E[l(g(X), Y)]$

Coming back to minimizing the probability of error: Let's say that we know all parameters of the distribution (X, Y), what is the optimal prediction rule which minimizes the probability of error (the risk)? This prediction rule is also called the **Bayes Predictor** g^* and its risk $R(g^*)$ is the **Bayes Risk**. We can show, that we minimize the probability of risk, if we classifier observations as belonging to class 1 if we know that their a posteriori probability of belonging to class 1 ($\eta(x)$) is bigger or equal to $\frac{1}{2}$.

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

We will now show, that this classification rule indeed leads to minimal risk among all classification rules g.

Theorem 1. *The Bayes classifier minimizes the probability of error. That is for any classifier g, $R(g) \geq R(g^*)$*

Proof. First we express the conditional probability of error for any classifier g.

$$\begin{aligned} P\left(g(x) \neq y | X = x\right) &= 1 - P\left(g(x) = 1, y = 1 | X = x\right) - P\left(g(x) = 0, y = 0 | X = x\right) \\ &= 1 - \mathbb{1}_{g(x)=1} P(y = 1 | X = x) - \mathbb{1}_{g(x)=0} P(y = 0 | X = x) \\ &= 1 - \mathbb{1}_{g(x)=1} \eta(x) - \mathbb{1}_{g(x)=0} (1 - \eta(x) \end{aligned}$$

We can see, that the probability of error depends on the a-posteriori probability $\eta(x)$. Now we compare the Bayes classifier g^* to any other classifier g.

$$\begin{aligned} &P(g(x) \neq y | X = x) - P(g^*(x) \neq y | X = x) \\ &\qquad = \eta(x) \left(\mathbb{1}_{g^*(x)=1} - \mathbb{1}_{g(x)=1}\right) + (1 - \eta(x)) \left(\mathbb{1}_{g^*(x)=0} - \mathbb{1}_{g(x)=0}\right) \\ &\qquad = (2\eta(x) - 1)(\left(\mathbb{1}_{g^*(x)=1} - \mathbb{1}_{g(x)=1}\right) \geq 0 \end{aligned}$$

Using the fact that $\mathbb{1}_{g(x)=0} = 1 - \mathbb{1}_{g(x)=1}$. By the defintion of g^*, this has to be bigger or equal to zero. Integrating both sides with respect to $\mu(dx)$ yields the theorem.

□

The proof shows, that the loss is

$$L(g) = 1 - \mathbb{E}\left[\mathbb{1}_{g(x)=1} \eta(x)\right] - \mathbb{E}\left[\mathbb{1}_{g(x)=0} (1 - \eta(x))\right]$$

In particular, the loss of the bayes classifier is

$$
\begin{aligned}
L(g^*) &= 1 - \mathbb{E}\left[\mathbb{1}_{g(x)>\frac{1}{2}}\eta(x)\right] - \mathbb{E}\left[\mathbb{1}_{g(x)\leq\frac{1}{2}}(1-\eta(x))\right] \\
&= \mathbb{E}\left[\min(\eta(X), 1-\eta(X))\right]
\end{aligned}
$$

If $\mathbb{1}_{g(x)>\frac{1}{2}}$ is true, then the Bayes risk is $1-\eta(x)$, and if $\mathbb{1}_{g(x)\leq\frac{1}{2}}$ then the Bayes risk is $\eta(x)$.

We can now also make statments about how hard a classification problem is. If the a posteriori probability $\eta(x)$ is close to $\frac{1}{2}$, then the two classes are almost inseperable and the problem is hard. On the other hand if $\eta(x)$ is 1 or 0, the classes are seperable. This doesn't necessarily mean that the problem will be easy to solve, since we generally have no information on $\eta(x)$. Typically, all we have is training data, which is modeled as as indepednednt pairs of random variables $(X_1, Y_1), ..., (X_n, Y_n)$, drawn from the same distribution as (X, Y). The training data set is also written as

$$
D_n((X_1, Y_1), ..., (X_n, Y_n))
$$

given the data, we construct a data-based classifier $g_n(x) = g_n(x, D_n)$. The classifier is random, as it depends on the random data. The probability of error

$$
R(g_n) = P\left(g_n(x) \neq y | D_n\right)
$$

is therefore a random variable. Clearly, $R(g_n) \geq R(g^*)$. $R(g_n) - R(g^*)$ is called the excess risk. We would like to construct classifiers, such that $R(g_n)$ is "small". The problem is, that we don't know anything about the distribution - though sometimes we may be willing to assume something about it. Further, what does it mean for a random variable to be small? We could possibly try to minimize $\mathbb{E}\left[R(g_n)\right] - R(g^*)$ or try to choose g so that $R(g_n) - R(g^*)$ is small with probability $1-\delta$. We will often relax our goal and fix a class $\mathbb{C}$ of classifiers g and try to minimize the risk over this particular class.

5. Nearest Neighbor Classification

A simple but effective classification rule (in terms of performance) is the nearest neighbor classifier. Let's assume that our set of observations χ lives in a metric space.