Open guide for Machine Learning: Theory

Roger Garriga Calleja

February 7, 2017

1 Important probability's inequalities

Markov's inequality

If $X \geqslant 0$ is a non-negative random variable and $t > 0$, then

$$P(X \geqslant t) \leqslant \frac{\mathbb{E}X}{t}. \tag{1}$$

Proof. $X \geqslant t\mathbb{1}_{X \geqslant t}$. Taking expectation on both sides: $\mathbb{E}X \geqslant \mathbb{E}\mathbb{1}_{X \geqslant t} = tP(X \geqslant t)$.
So, $P(X \geqslant t) = \frac{\mathbb{E}\mathbb{1}_{X \geqslant t}}{t}$.
Q.E.D □

Chebyshev's inequality

For any random variable with finite variance

$$P(|X - \mathbb{E}X| \geqslant t) \leqslant \frac{\text{Var}X}{t^2}. \tag{2}$$

Proof. $P(|X - \mathbb{E}| \geqslant t) = P((X - \mathbb{E})^2 \geqslant t^2)$. By Markov's inequality $P((X - \mathbb{E})^2 \geqslant t^2) \leqslant \frac{\mathbb{E}(X-\mathbb{E})^2}{t^2} = \frac{\text{Var}X}{t^2}$.
Q.E.D □

In particular, if $S = \sum_{i=1}^{n} x_i$ independent, then $\text{Var}(S) = \sum_{i=1}^{n} \text{Var}(x_i)$ and $= n\text{Var}(x_1)$ if they are independent identically distributed (iid from now on). So,

$$P(|S - \mathbb{E}S| \geqslant t) \leqslant \frac{n\text{Var}(x_1)}{t^2}. \tag{3}$$

That implies the weak law of larges numbers, dividing by n:

$$P(|\frac{1}{n}S - \frac{1}{n}\mathbb{E}S| \geqslant \epsilon) \leqslant \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty. \tag{4}$$

Chernoff's bounds

For any $\lambda > 0$,

$$P(X - \mathbb{E}X \geqslant t) \leqslant \frac{\mathbb{E}e^{\lambda(X - \mathbb{E}X)}}{e^{\lambda t}}. \tag{5}$$

Proof. $P(X - \mathbb{E}X \geqslant t) = P(e^{\lambda(X-\mathbb{E}X)} \geqslant e^{\lambda t})$, now applying Markov's inequality $P(e^{\lambda(X-\mathbb{E}X)} \geqslant e^{\lambda t}) \leqslant \frac{\mathbb{E}e^{\lambda(X-\mathbb{E}X)}}{e^{\lambda t}}$.
Q.E.D □

Hoeffding's inequality

Hoeffding's Lemma: If X is a random variable taking values in $[a, b]$, then $\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leqslant e^{\frac{\lambda^2(b-a)}{8}}$. (In particular for $X \in [0,1]$, $\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leqslant e^{\frac{\lambda^2}{8}}$).

Though Hoeffding's Lemma and Chernoff's bounds, we get the Hoeffding's inequality:

$$P(S - \mathbb{E}S \geqslant t) \leqslant e^{-\frac{2t^2}{n(b-a)^2}}. \tag{6}$$

(In particular for $X \in [0,1]$, $P(S - \mathbb{E}S \geqslant t) \leqslant e^{-\frac{2t^2}{n}}$)

Proof. Let S be the sum of n iid random variables, by Chernoff's bounds $P(S-\mathbb{E}S \geqslant t) \leqslant \min\limits_{\lambda>0} \frac{\prod_{i=1}^{n} \mathbb{E}[e^{\lambda(X_i-\mathbb{E}X_i)}]}{e^{\lambda t}}$, now using Hoeffding's Lemma, $\min\limits_{\lambda>0} \frac{\prod_{i=1}^{n} \mathbb{E}[e^{\lambda(X_i-\mathbb{E}X_i)}]}{e^{\lambda t}} \leqslant \min\limits_{\lambda>0} \frac{e^{\lambda^2 n(b-a)}}{\lambda t} = \min\limits_{\lambda>0} e^{\frac{\lambda^2 n(b-a)}{8} - \lambda t}$. Minimizing (taking derivative to 0) we get $\lambda = \frac{4t}{n(b-a)}$, so $P(S - \mathbb{E}S \geqslant t) \leqslant e^{-\frac{2t^2}{n(b-a)^2}}$
Q.E.D □

Normalizing we get $P(\frac{1}{\sqrt{n}}(S - \mathbb{E}S) \geqslant t) \leqslant e^{-\frac{2t^2}{(b-a)^2}}$.

Bernstein's inequality

Let $X_1, \ldots, X_n$ be independent such that $X_i \leqslant 1$ $\forall i$ and let $v = \sum\limits_{i=1}^{n} \mathbb{E}[X_i^2]$. Then, $\forall t > 0$,

$$P(\sum_{i=1}^{n} \geqslant \mathbb{E}\sum_{i=1}^{n} X_i + t) \leqslant e^{-\frac{t^2}{2(v+\frac{t}{3})}}. \tag{7}$$

If X_i are iid with $\mathbb{E}X = 0$, then $v = n\sigma^2$ ($\sigma^2 = \mathrm{Var}X$), so

$$P(\sum_{i=1}^{n} X_i \geqslant t) \leqslant e^{-\frac{t}{2n\sigma^2 + \frac{2}{3}t}}. \tag{8}$$

2 Mean estimator

The motivation of this section is to find a good estimator of the expected value of a variable X given n observations of the variable. That said, we assume that $x_1, \ldots, x_n$ are independent identically distributed (iid from now on) random variables with expected value $\mathbb{E}X = m$.

The estimator will be a function $\hat{m}_n(x_1, \ldots, x_n)$ of the observations we have. A good estimator should have "small" error $|\hat{m}_n - m|$. However, since $\hat{m}_n$ is a random variable (it is a function of random variables) there are many ways to measure the error. In general it is measured as the expected value of a function $l : \Re \times \Re \to \Re_+$ called loss function, that symbolizes how much we "pay" by saying $m = \hat{m}_n$. Common examples of loss functions are $l(\hat{m}_n, m) = (\hat{m}_n - m)^2$ and $l(\hat{m}_n, m) = |\hat{m}_n - m|$. A more flexible way of measuring the error is using the probability that $\hat{m}_n$ is at distance more than ϵ, $P(|\hat{m}_n - m| > \epsilon) = \mathbb{E}\mathbb{1}_{|\hat{m}_n - m| > \epsilon}$. This corresponds to the loss function $l(\hat{m}_n, m) = \mathbb{1}_{|\hat{m}_n - m| > \epsilon}$.

The naive estimator is the sample mean $\hat{m}_n = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$, which is unbiased and has a mean squared error (MSE) of $\mathbb{E}(\hat{m}_n - m) = \frac{\sigma^2}{n}$. But it behaves poorly in general if the variance is large. The probability of being far from the real mean can be bounded using the inequalities:

- By Chebyshev's: $P(|\hat{m}_n - m| \geqslant \epsilon) \leqslant \frac{\sigma^2}{n\epsilon^2}$.
- By Hoeffding's: If $X_i \in [0,1]$, then $P(|\hat{m}_n - m| \geqslant \epsilon) \leqslant e^{-2n\epsilon^2}$.
- Applying Markov's in the sub-gaussian case ($\mathbb{E}e^{\lambda(X\mathbb{X})} \leqslant e^{\frac{\lambda^2\sigma^2}{2}}$) like we did in Hoeffding's lemma : $P(|\hat{m}_n - m| \geqslant \epsilon) \leqslant e^{-\frac{n\epsilon^2}{2\sigma^2}}$.

In general it is difficult that we can apply Hoeffding's inequality or that we have a sub-gaussian distribution (in order to have $\mathbb{E}e^{\lambda X} = \int e^{\lambda x} f(x)\mathrm{d}x < \infty$ we need the density $f(x)$ to beat $e^{\lambda x}$). So we should find an estimator of the mean more stable than the sample mean.

2.1 Median of means estimator (MoM)

The idea behind this estimator is to divide the data into K blocks of size $l = \frac{n}{K}$ each, compute the mean in each block and compute the median of the means.

So, the K blocks would be $\{x_1, \dots, x_l\}, \{x_{l+1}, \dots, x_{2l}\}, \dots \{x_{(K-1)l+1}, \dots, x_{Kl}\}$, the means $\mu_1 = \frac{1}{l}\sum\limits_{i=1}^{l} x_i, \mu_2 = \frac{1}{l}\sum\limits_{i=l+1}^{2l} x_1, \dots, \mu_K = \sum\limits_{i=(K-1)l+1}^{Kl} x_i$ and the estimator $\hat{m}_n = \text{median}(\mu_1, \dots, \mu_k)$.

Assuming that $\text{Var}X = \sigma^2 < \infty$, by Chebyshev

$$|\mu_i - m| < \frac{2\sigma}{\sqrt{l}} \text{ with probability } \geqslant \frac{3}{4}, \tag{9}$$

for each $j = 1, \dots, K$ (we could took a probability different of $\frac{3}{4}$ that may result in a better constant, but $\frac{3}{4}$ is good enough).

Proof. $P(|X - \mathbb{E}X| \geqslant \epsilon) \leqslant \frac{\sigma^2}{n\epsilon^2} = \delta \Leftrightarrow$ with probability $\geqslant 1 - \delta$, $|X - \mathbb{E}X| < \frac{\sigma}{\sqrt{n\delta}}$. Taking $\delta = \frac{1}{4}$, $\frac{\sigma}{\sqrt{n\delta}} = \frac{2\sigma}{\sqrt{n}}$. □

And for the estimator, $|\hat{m}_n - m| \geqslant \frac{2\sigma}{\sqrt{l}}$ iif at least half of the $\mu_1, \dots, \mu_K$ are $\frac{2\sigma}{\sqrt{l}}$ away from m. Then, the problem is reduced to the binomial and Hoeffding's inequality can be applied:

$$P(\text{Bin}(K, \frac{1}{4}) \geqslant \frac{K}{2}) = P(\text{Bin}(K, \frac{1}{4} - \frac{1}{4}) \geqslant \frac{K}{4}) \leqslant e^{-\frac{2K^2}{16K}} = e^{-\frac{K}{8}}. \tag{10}$$

Then, we can choose K according to the precision δ we want, $e^{-\frac{K}{8}} = \delta \Rightarrow K = [8\log\frac{1}{\delta})]$ so $l = \frac{n}{8\log\frac{1}{\delta}}$.

Result: MoM estimator with parameter $K = [8\log\frac{1}{\delta}]$ satisfies that $|\hat{m}_n - m| \leqslant 2\sigma\sqrt{\frac{8\log\frac{1}{\delta}}{n}}$. Notice that this inequality is sub-gaussian,

$$P(|\hat{m}_n - m| \geqslant \epsilon) \leqslant e^{-\frac{n\epsilon^2}{2\sigma^2}} = \delta \Leftrightarrow |\hat{m}_n - m| < \sigma\sqrt{\frac{2\log\frac{1}{\delta}}{n}} \text{ with probability } 1 - \delta, \tag{11}$$

This bound is much better than the one obtained by Chebyshev. However it has two downsides: MoM is not unbiased and the estimator depends on the precision δ.

3 Dimensionality Reduction

3.1 By Random Projection

The idea is to map a set of $a_1, a_2, \ldots, a_N$ points that belong to a space of dimension D, $\Re^D$ into a smaller space of dimension d, $\Re^d$ with $d << D$. So we look for a function $f : \Re^D \to \Re^d$ such that $f(a_1), \ldots, f(a_N)$ is a good representation of $a_1, \ldots, a_N$.

In this case the fundamental characteristic of "good representation" would be that the pairwise distances between points are preserved. This characteristic is important on clustering. But in general, if inside the set $\{a_1, \ldots, a_N\}$ there are s points that are independent, then at most we can only reduce the dimension to $\Re s - 1$, which is not enough because we could have a number of points of the order of N, so the dimensionality reduction would be minimal.

However, if we allow some slack, the situation changes dramatically. We require that f is such that

$$1 - \epsilon \leqslant \frac{||f(a_i) - f(a_j)||^2}{||a_i - a_j||^2} \leqslant 1 + \epsilon, \text{ for some } \epsilon > 0. \tag{12}$$

And it turns out that such an f exists whenever $d \geqslant 8\frac{\log N}{\epsilon^2}$ (Johnson-Lindenstrauss lemma). But, even though we have existence, we do not know how to construct it. The good thing is that most linear functions would work, so we can pick an f randomly.

As f is a projection from $\Re^D$ space to $\Re^d$, it has a projection matrix W associated and then $f(a) = Wa$, where $W = (W_{ij})$ is a $d \times D$ matrix. We take $W_{ij} \sim N(0, \frac{1}{d})$.
We want to see that $||W(a_i - a_j)||^2 \approx ||a_i - a_j||^2$. For any fixed $b \in \Re D$,

$$\mathbb{E}||Wb||^2 = \mathbb{E}\sum_{i=1}^{d}\left(\sum_{j=1}^{D} W_{ij}b_j\right)^2 = \sum_{i=1}^{d}\mathbb{E}\left(\sum_{j=1}^{D} W_{ij}b_j\right)^2 \underset{\text{ind}}{=} \sum_{i=1}^{d}\sum_{j=1}^{D}\left(\mathbb{E}W_{ij}^2 bj^2\right) \tag{13}$$

$$\underset{b_j \text{ ctt}}{=} \sum_{i=1}^{d}\sum_{j=1}^{D} b_j \mathbb{E}W_{ij}^2 = \sum_{i=1}^{d}\frac{1}{d}\sum_{j=1}^{D} b_j^2 = ||b||^2. \tag{14}$$

In particular, $\mathbb{E}||W(a_i - a_j)||^2 = ||(a_i - a_j)||^2$. Now we want it with high probability,

$$\left|\frac{W(a_i - a_j)||^2}{||(a_i - a_j)||^2} - 1\right| < \epsilon. \tag{15}$$

We will denote $c_{ij} = \frac{(a_i - a_j)}{||a_i - a_j||}$, (observe that c_{ij} is a unit vector).

Then, $\|Wc\|^2 - 1 = \sum_{i=1}^{d}\left(\sum_{j=1}^{D} W_{ij}c_j\right)^2 - 1$, as c is unit vector and W_{ij} are $N(0, \frac{1}{d})$, $\sum_{j=1}^{D} W_{ij}c_j \sim N(0, \frac{1}{d})$.
Also, $\sum i = 1^d \mathbb{E}[N_i^2] = 1$. Putting both things together, we can write

$$\|Wc\|^2 - 1 = \sum_{i=1}^{d}\left(N_i^2 - \mathbb{E}[N_i^2]\right). \tag{16}$$

Now, we can apply Chernoff bounds $P(|\|Wc_{ij}\| - 1 > \epsilon) \leqslant e^{-\frac{\epsilon^2 d}{4}}$, which implies

$$P(\max_{i,j=1,\ldots N}\|Wc_{ij}\|^2 - 1 > \epsilon) \leqslant \binom{N}{2} e^{-\frac{\epsilon^2 d}{4}} \leqslant \frac{N^2}{2}e^{-\frac{\epsilon^2 d}{4}}. \tag{17}$$

If we want the probability to be less than δ, then we find that $d \geqslant \frac{4}{\epsilon^2} \log \frac{N^2}{2\delta}$

Johnson-Lindestrauss lemma: Given $a_1, \ldots, a_n$, there exists $f : \Re^D \to \Re^d$ such that for all $i, j = 1, \ldots, N$,

$$1 - \epsilon \leqslant \frac{\|f(a_i) - f(a_j)\|^2}{\|a_i - a_j\|^2} \leqslant 1 + \epsilon. \tag{18}$$

whenever $d \geqslant \frac{8}{\epsilon^2} \log N$.

Especial remarks after this result is that the best dimension we can hope for (even allowing f to be non-linear) is $d \sim \frac{\log n}{\epsilon^2}$. That no deterministic construction is known. And that Euclidian distance is crucial in dimensionality reduction.

4 Basic decision theory

The idea is to predict the value of a random variable $Y \in T$ assuming that we know the distribution of Y. The set T may be $\{0, 1\}, \{1, \ldots, N\}, \Re, \Re^d, \ldots$. We will note our prediction as $p \in T$. The quality of the predictor is measured by a lost function

$$l : T \times T \to \Re_+.$$

There are many functions that can be used as loss functions. For $T = \{0, 1\}$ a natural one is $l(y, p) = \mathbb{1}_{y \neq p}$. An asymmetric version that penalize more one error than the other would be

$$l(y, p) = \begin{cases} a & \text{if } y = 0, p = 1 \\ b & \text{if } y = 1, p = 0 \\ 0 & \text{otherwise} \end{cases}$$

For $T = \Re^d$ we may have

$$l(y, p) = \|y - p\|^2,$$
$$l(y, p) = \|y - p\|.$$

We will focus on the *risk* of the decision p. Which is

$$R(p) = \mathbb{E}l(y, p). \tag{19}$$

For the risk function, the optimal prediction would be $p^* \in T$ such that $p^* = \underset{p \in T}{\text{argmin}} R(p)$. Then, the optimal risk is noted $R^* = R(p^*)$.

For example, if $T = \{0, 1\}$ and $l(y, p) = \mathbb{1}_{y \neq p}$, the risk would be

$$R(p) = \mathbb{E}l(Y, p) = p(Y = 0)\mathbb{1}_{p=1} + p(Y = 1)\mathbb{1}_{p=0}. \tag{20}$$

Then, $R^* = R(p^*) = \min(p(Y = 0), p(Y = 1))$.

For $T = \Re$ and $l(y, p) = (y - p)^2$, the risk would be $R(p) = \mathbb{E}(Y - p)^2 = \mathbb{E}[Y^2] - 2p\mathbb{E}[Y] + p^2$. The minimizer would be $p^* = \mathbb{E}Y$. If the loss function was $l(y, p) = |y - p|$, p^* would be the median of Y. That is also a way to define the median for higher dimensional spaces $Med(Y) = \underset{p}{\text{argmin}} \|Y - p\|^2$.

4.1 Prediction problem with a covariate

Let (X, Y) be a pair of random variables taking values in $\Omega \times T$. We usually call X observations, features or covariates and Y label or target. We want to predict Y given an observed value X. If we assume that the joint distribution of (X, Y) is known, a predictor p is a function of X $p : \Omega \times T$. Given a loss function $l : T \times T \to \Re_+$, the loss function will be $l(p(X), Y)$ and the risk $R(p) = \mathbb{E}l(p(X, Y))$. Then, the risk can also be written as

$$R(p) = \mathbb{E}[l(p(x), Y)] = \mathbb{E}_X \mathbb{E}_{Y|X}[l(p(X), Y)|X]. \tag{21}$$

Example
Binary classification: $T = \{0, 1\}$, $l(p, Y) = \mathbb{1}_{p \neq Y}$. Then the risk function would be $R(p) = \mathbb{E}\mathbb{1}_{p(X) \neq Y} = P(p(X \neq Y)) = \mathbb{E}\mathbb{P}(p(X) \neq Y|X)$. So p would be minimized as $p^* = \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$. This $p(Y = 1|X)$ is called a posteriori probability (sometimes noted by $\eta(X)$). Then, the optimal risk would be $\mathbb{E}[\min(\eta(X), 1 - \eta(X))]$.
Basically the risk minimization and optimal predictor with a covariate work as without it but instead of marginal probabilities we take conditional probabilities $Y|X$.

Notation in binary classification:
A posteriori probability: $\eta(x) = \mathbb{P}(Y = 1|X = x) = \mathbb{E}(Y|X = x)$.
A priori probabilities: $q_1 = \mathbb{P}(Y = 0)$, $q_1 = \mathbb{P}(Y = 1)$.
Marginal distribution of X: $\mu(A) = \mathbb{P}(X \in A)$.
Conditional densities for X given Y: $f_0(X) = f(X|Y = 0)$, $f_1 = (X|Y = 1)$.

Example

Let $f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \sim N(0, 1)$, $f_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2}}$, $q_0 = \frac{1}{2}$, $q_1 = \frac{1}{2}$. Then the a posteriori probabilities would be

$$\mathbb{P}(Y = 1|X = x) = \frac{f_1(x)q_1}{f_1(x)q_1 + f_0(x)q_0},$$

$$\mathbb{P}(Y = 0|X = x) = \frac{f_0(x)q_0}{f_1(x)q_1 + f_0(x)q_0}.$$

Then, $p^*(x) = \begin{cases} 1 & \text{if } \frac{f_1(x)}{f_0(x)} > \frac{q_0}{q_1} \\ 0 & \text{otherwise} \end{cases}$. If $q_0 = q_1 = \frac{1}{2}$ we choose the i with $f_i(x)$ higher, otherwise it would depend on how different are the marginals.

The problem is that in general we do not know the conditional distribution $X|Y$, we have data instead $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. Our assumption will be that the (X_i, Y_i) are iid. With that, we will try to predict the next Y_{n+1} given a new observation X_{n+1}. So, the predictor will depend on the new observation and the data $p(X_{n+1}, D_n) = p_n(X_{n+1})$. The goal is to find a predictor p_n such that the risk $R(p_n) = \mathbb{E}[l(p_n(X_{n+1}), Y)|D_n]$ is as close to R^* as possible.
We will measure how far is the predictor's risk from R^* ($\mathbb{P}(R(p_n) - R^* > \epsilon)$) and minimize it for all distributions.

We define a prediction rule to be **consistent** if

$$\mathbb{E}R(p_n) \underset{n \to \infty}{\longrightarrow} R^* \Leftrightarrow \mathbb{P}(R(p_n) - R^* > 0) \to 0 \ \forall \epsilon > 0.$$

No-free-lunch theorem: For all prediction rules p_n and every sample size n, no matter how big n is, there exist a distribution of X, Y such that $\mathbb{E}R(p_n) - R^* > 0.49$ (for classification, but in general is the same, there always exist a distribution that makes the prediction perform bad).

In order to remedy this, we restrict the class of distributions we target or the class of prediction rules. So R^* will be the minimum in the class. The next step is to estimate the risk using a good estimator $\hat{R}_n(p_n)$ of $R(p_n)$. This estimate should be such that $|R(p_n) - \hat{R}_n(p_n)|$ is small for all distribution given a sample of size n. At least $R(p_n) \leqslant \hat{R}(p_n)$ and small with high probability.

4.2 Classification

First suppose that our data is $D_n = (Y_1, \dots, Y_n)$, $Y_i \in \{0,1\}$ (no covariates X).

As a first approach we could consider $p_n = Y_1$, $R(p_n) = \mathbb{1}_{Y_1=0}\mathbb{P}(Y=1) + \mathbb{1}_{Y=1}\mathbb{P}(Y=0)$. However, as $R^* = \min\{\mathbb{P}(Y=0), \mathbb{P}(Y=0)\}$, the expected value will not be consistent $R^* \leqslant \mathbb{E}R(p_n) = 2\mathbb{P}(Y=0)\mathbb{P}(Y=1) \leqslant 2R^*$.

A better idea is to pick p_n as the majority rate, meaning $p_n = \begin{cases} 0 & \text{if } \frac{1}{n}\sum\limits_{i=1}^{n} Y_i < \frac{1}{2} \\ 1 & \text{otherwise} \end{cases}$. For n big that would tend to $p^* = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$.
This predictor is consistent: $R(p_n) = \eta\mathbb{1}_{\sum_{i=1}^{n} Y_i < \frac{n}{2}} + (1-\eta)\mathbb{1}_{\sum_{i=1}^{n} Y_i > \frac{n}{2}}$. The expected value of the risk would be

$$\mathbb{E}R(p_n) = \eta\mathbb{P}(\text{Bin}(n,\eta) < \frac{n}{2}) + (1-\eta)\mathbb{P}(\text{Bin}(n,\eta) \geqslant \frac{n}{2}). \tag{22}$$

Using that $\mathbb{P}(\text{Bin}(n,\eta) \geqslant \frac{n}{2}) = 1 - \mathbb{P}(\text{Bin}(n,\eta) < \frac{n}{2})$ and assuming without lost of generality $\eta < \frac{1}{2}$ we get $\mathbb{E}R(p_n) = \eta + (1-2\eta)\mathbb{P}(\text{Bin}(n,\eta) \geqslant \frac{n}{2})$. Since in this case $R^* = \eta$, using Hoeffding's inequality we get that

$$\mathbb{E}R(p_n) - R^* = \eta + (1-2\eta)\mathbb{P}(\text{Bin}(n,\eta) \geqslant \frac{n}{2}) \leqslant e^{-2n(\frac{1}{2}-\eta)^2}, \tag{23}$$

which decreases fast with n. The maximum will be taken for $\eta \approx \frac{1}{2} - \frac{1}{\sqrt{n}}$.
Now let's consider a more realistic situation in which we have also covariates. Our data would be $D_n = (X_1, Y_1), \dots, (X_n, Y_n)$, where $(X_n, Y_n) \in \Omega \times \{0,1\}$. In this framework the goal is to predict Y_{n+1} once we have observed X_n. The optimal decision would be

$$p^*(X) = \begin{cases} 1 & \text{if } \eta(X) \geqslant \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}, \tag{24}$$

known as Bayes decision (even though Bayesian Statistics is not involved) (recall $\eta(x) = \mathbb{P}(Y=1|X=x)$). Then, the optimal risk (called Bayes risk) would be

$$R^* = \mathbb{E}\min\{\eta(X), 1-\eta(X)\}. \tag{25}$$

4.2.1 The nearest neighbor rule

With this rule, we predict the new observation as the same category as his nearest neighbor. We would order the data as $(X_{(1)}(x), Y_{(1)}(x), \dots, (X_{(n)}(x), Y_{(n)}(x)$ according to the distance to X. That is $d(X_{(1)}(x), x) \leqslant \cdots \leqslant d(X_{(n)}(x), x)$. The first question is how to define the notion of distance in each problem. We could use $d(x,y) = \|x-y\|_1, \dots, d(x,y) = \|x-y\|_\infty$, or also in general

$d(x,y) = \sqrt{(x-y)^T A(x-y)}$, where A is a symmetric positive semidefinite matrix.

However, when we increase the dimension we suffer the **curse of dimensionality**, the probability of being near decays exponentially with the dimensions. Suppose that our X are uniformly distributed in $[0,1]$, then

$$\mathbb{P}(d(x, X_{(1)}(x)) > \epsilon) = \mathbb{P}(d(x,X_1) > \epsilon), \ldots, \mathbb{P}(d(x,X_n) > \epsilon) = \mathbb{P}(d(x,X_1) > \epsilon)^n = (1 - \mathbb{P}(d(x,X_1) \leqslant \epsilon)). \tag{26}$$

As X are iid uniform distributions, $\mathbb{P}(d(x,X_1) \leqslant \epsilon) = \epsilon^d v_d$, where v_d is the volume of the unit ball in $\Re^d$. So

$$\mathbb{P}(d(x, X_{(1)}(x)) > \epsilon) = (1 - v_d \epsilon^d)^n \leqslant e^{-v_d n \epsilon^d}, \tag{27}$$

(using $1 - x \approx e^{-x}$ for x small). That will be small if $n\epsilon^d$ is huge, meaning that $\epsilon \geqslant \sqrt[d]{n}$. That means that for $d = 1$, $d(x, X_{(1)}(x))$ is of the order $\frac{1}{n} < \frac{1}{100} \Leftrightarrow n \geqslant 100$; for $d = 2$, $\frac{1}{n} < \frac{1}{100} \Leftrightarrow n \geqslant 10000$; for $d = 10$, $\frac{1}{n} < \frac{1}{100} \Leftrightarrow n \geqslant 100^{10}$. So the distance between points increases very fast with the dimension. However, $\forall d$, $d(x, X_{(1)}(x)) \to 0$ in probability when $n \to \infty$. This is true for all distributions.

Remark: Nearest neighbor works if the classes are well separated. Then, even if the distance is big, it is still much smaller than the distance to the nearest of the other class. This tells us that in order to increase dimensionality (adding new variables) we better do it with variables that give us real information. Adding unimportant information adds noise that difficult classification.

Theorem: The expected risk of the nearest neighbor rule is

$$\lim_{n\to\infty} \mathbb{E}R(p_n^{NN}) = 2\mathbb{E}[\eta(x)(1-\eta(x))] = R_{NN}. \tag{28}$$

That is true for all distributions.

As $\min(\eta, 1-\eta) \leqslant 2\eta(1-\eta) \leqslant 2\min(\eta, 1-\eta)$, we always have that $R^* \leqslant R_{NN} \leqslant 2R^*(1-R^*)$, so R_{NN} will only be consistent whenever $R^* = 0$. That means that it will only be consistent if we Y is a deterministic function of X.

4.2.2 K-nearest neighbor rule

The K-NN rule takes a majority vote among the K closest points to predict. This rule have the same property as the previous one, $d(X_{(k)}(x), x) \underset{n\to\infty}{\longrightarrow} 0$.

The risk for a general K would be $R^{K-NN} = P(Y \neq \text{majority}\{Y_{(1)}, \ldots, Y_{(K)}\})$. For $K = 3$ this will happen if $Y = 1$ and the at least two of the nearest neighbors are 0, that is if either two of them are 0 or the three are 0 (similarly for $Y = 0$). So, $P(Y \neq \text{majority}\{Y_{(1)}, Y_{(3)}, Y_{(3)}\}) = \eta[(1-\eta)^3 + 3\eta(1-\eta)^2] + (1-\eta)[\eta^3 + 3(1-\eta)\eta^2]$ working out the equations we get $P(Y \neq \text{majority}\{Y_{(1)}, Y_{(3)}, Y_{(3)}\}) = \eta(1-\eta) + 4\eta^2(1-\eta)^2$. Then, the risk of the 3 nearest neighbor will be

$$R^{3-NN} = \mathbb{E}[\eta(x)(1-\eta(x))] + 4\mathbb{E}[\eta(x)^2(1-\eta(x))^2] \leqslant 2\mathbb{E}[\eta(x)(1-\eta(x))] = R^{NN} \tag{29}$$

If we develop in general for any K, we get

$$R^{K-NN} =\mathbb{E}[\eta(X)\mathbb{P}\left(\text{Bin}(K,\eta(X)) < \frac{K}{2}|X\right) + (1-\eta)\mathbb{P}\left(\text{Bin}(K,\eta(X)) > \frac{K}{2}|X\right)] = \tag{30}$$

$$=\mathbb{E}[\min(\eta(X), 1-\eta(X))] + \mathbb{E}[(2\eta(X)-1)\mathbb{P}\left(\text{Bin}(K, \min(\eta, \eta-1)) > \frac{K}{2}|X\right)] \leqslant \tag{31}$$

$$\underset{\text{Hoeffding}}{\leqslant} R^* + e^{-2K|\eta(X)-\frac{1}{2}|^2} \leqslant R^* + \max_{\eta}|2\eta-1|e^{-2K|\eta(X)-\frac{1}{2}|^2} \leqslant \tag{32}$$

$$\leqslant R^* + \frac{1}{\sqrt{eK}}, \tag{33}$$

where the last inequality comes by maximizing the expression. This bound decreases with the number of neighbors we take, if K goes to infinity in a way that $\frac{K}{n}$ goes to 0 (for example $K = \sqrt{n}$ or $K = \log n$) then the risk converges to R^* for all distributions. So, it would be universally consistent.

Nearest neighbor rules are examples of averaging rules, the simplest such rules are partitioning classifiers. These are rules that partition the space into small cubes and take a majority vote within the cell of the partition where the point X to be classified lies. If h is the size of the cube, and $h \to 0$ such that $nh^d \to \infty$ for $n \to \infty$, then the rule is universally consistent. Although the problem becomes dangerous when we increase the dimensions, because when one divides the space in cells of length h the number of cells increases exponentially with the dimension. That makes the space full of cells and, as in order to be able to average inside each cell we need a fair number of points on it, we would need a lot of points to populate all the cells. We suffer the curse of dimensionality.

4.3 Empirical risk minimization

Let $g : \Omega \to \{0,1\}$ be a classifier that not depends on data. Given data $D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$, to know how good the estimator is we can estimate the risk as $R(g) = \mathbb{P}(g(X) \neq Y)$ by $R_n(g) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g(X_i)\neq Y_i}$. By Hoeffding's inequality we know that with probability $\geqslant 1-\delta$,

$$|R_n(g) - R(g)| \leqslant \sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \tag{34}$$

Now, the idea would be to know among different classifiers which is the best. That is given N classifiers $g_1, \ldots, g_N : X \to \{0,1\}$, choose the one that has less risk. We can estimate the risk with the same data and pick the best one, but this does not mean that in general it will be the best classifier.

We can estimate the risk of each classifier g_j as $R_n(g_j) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{g_j(X_i)\neq Y_i}$. For each classifier we have the above inequality, but the risks are not independent of the classifier rule used. We will use the union bound ($\mathbb{P}(\bigcup_{i=1}^{n} A_i) \leqslant \sum_{i=1}^{n} \mathbb{P}(A_i)$).

$$\mathbb{P}(\max_{j=1,\ldots,N}|R_n(g_j) - R(g_j)| \geqslant \epsilon) = \mathbb{P}(\bigcup_{j=1}^{n}\{|R_n(g_j) - R(g_j)| \geqslant \epsilon\}) \tag{35}$$

$$\leqslant \sum_{j=1}^{n} \mathbb{P}(|R_n(g_j) - R(g_j)| \geqslant \epsilon) \leqslant 2Ne^{-2n\epsilon^2} = \delta. \tag{36}$$

Now, solving for δ we get that with probability $\geqslant 1-\delta$,

$$|R_n(g_j) - R(g_j)| \leqslant \sqrt{\frac{\log \frac{2N}{\delta}}{2n}}. \tag{37}$$

One important consequence of this is that if we take a classifier rule depending on data D_n $\hat{p}_n : \Omega \times (\Omega \times \{0,1\})^n \to \{0,1\}$ that picks a classifier from the set $\{g_1, \dots, g_N\}$. And we estimate the risk of $\hat{p}_n$ by $R_n(\hat{p}_n) = \frac{1}{n}\mathbb{1}_{\hat{p}_n(X_i)\neq Y_i}$, then with probability $\geqslant 1-\delta$

$$|R_n(\hat{p}_n - R(\hat{p}_n))| \leqslant \sqrt{\frac{\log \frac{2N}{\delta}}{2n}}. \tag{38}$$

Which implies that we can bound the theoretical risk of $\hat{p}_n$ from data as $R(\hat{p}_n) \leqslant R_n(\hat{p}_n) + \sqrt{\frac{\log \frac{2N}{\delta}}{2n}}$ (with probability $\geqslant 1-\delta$). We still have the bound depending on the number of classifiers in the class, which can be huge.

Now, we will denote $\hat{p}_n = \underset{g_1,\dots,g_N}{\operatorname{argmin}} R_n(g_j)$ the chosen classifier that empirically minimizes the risk and $\bar{p} = \underset{g_1,\dots,g_N}{\operatorname{argmin}} R(g_j)$ the best classifier (that minimizes the theoretical risk). Let's see how close they are regarding their risk bounding the excess of risk:

$$R(\hat{p}_n) - R(\bar{p}) = (R(\hat{p}_n) - R_n(\hat{p}_n)) + (R_n(\hat{p}_n) - R(\bar{p}_n)). \tag{39}$$

Observe that $R(\hat{p}_n) - R_n(\hat{p}_n) \leqslant \max_{j=1,\dots,N} R(g_j) - R_n(g_j)$ ($\hat{p}_n$ is one of the g_j), also $R_n(\hat{p}_n) \leqslant R_n(\bar{p}_n)$ ($\hat{p}_n$ is chosen minimizing the empirical risk and $\bar{p}_n$ is the theoretical minimum) $\Rightarrow R_n(\hat{p}_n) - R(\bar{p}_n) \leqslant \max_{j=1,\dots,N} R_n(g_j) - R(g_j)$. Using (38) we can bound the excess risk by $R(\hat{p}_n) - R(\bar{p}) \leqslant \sqrt{\frac{2\log \frac{2N}{\delta}}{n}}$ with probability $\geqslant 1-\delta$. This bound is very conservative, working a bit more we can get to

$$R(\hat{p}_n) - R(\bar{p}) \leqslant c\left(\sqrt{\frac{R(\bar{p})\log \frac{N}{\delta}}{n}} + \frac{\log \frac{N}{\delta}}{n}\right),\ c \text{ constant}. \tag{40}$$

The resubstitution estimate:

Let p_n be a data-based estimate trained on $D_n = (X_1, Y_1), \dots, (X_n, Y_n)$. We want to estimate the risk $R(p_n) = \mathbb{P}(p_n(X) \neq Y | D_n)$ based on the same data D_n. The resubstitution estimate is just counting how many errors we did on our data

$$R_n(p_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{p_n(x_i)\neq Y_i}. \tag{41}$$

Observe that in this case this is not a sum of independent variables, because p_n depends on X_i. However, if p_n is given by a fixed class $\mathcal{C}$ of non-data-dependent classifiers, then

$$|R_n(p_n) - R(p_n)| \leqslant \max_{g\in\mathcal{C}} |R_n(g) - R(g)|. \tag{42}$$

If the classifier chooses from a small class, then the resubstitution estimate would work because we could bound it, otherwise it would be impossible. Observe that in particular the 1-NN has $R_n(p_n) = 0$

(for each point the nearest neighbor will be itself).

The deleted (or leave one out crossvalidation) estimate

In this case the empirical risk would be computed as the sum of errors made on each point, but for every point i we would train a $p_n^{(i)}$ with the data except i, so $p_n^{(i)}$ would be train on $D_n^{(-i)} = (X_1, Y_1), \dots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \dots, (X_n, Y_n)$.

$$R_n^{(D)}(p_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{p_n^{(i)}(X_i)\neq Y_i}. \tag{43}$$

Now, we don't have the dependence problem, $p_n^{(}i)$ does not depend on (X_i, Y_i). And

$$\mathbb{P}(p_n^{(i)}(X_i) \neq Y_i | D_n^{(} - i)) = R(p_n^{(}i)), \tag{44}$$

Taking expectation over the X_i we get

$$\mathbb{E}R_{(}p_n^{(i)}) = \mathbb{E}R(p_n^{(}j)),\ \text{for any } j = 1, \dots, n. \tag{45}$$

In particular

$$\mathbb{E}R_{(}p_n^{(i)}) = \mathbb{E}R(p_n^{(}n)) = \mathbb{E}R(p_{n-1}). \tag{46}$$

So, the expected value is almost unbiased,

$$\mathbb{E}R_n^{(}D)(p_n) = \mathbb{E}R(p_{n-1}). \tag{47}$$

In general saying something about $\mathbb{P}(|R_n^{(D)} - R(p_n)| > t|)$ is a challenge. All the dependencies that are when leaving one out are not trivial, so a lot of different classes have been studied to see in their particular case how it behaves. In practice we use the approximation that it is the same more or less when leaving one out, which works fairly well if we have a lot of data.

Data splitting

Suppose we have N data-dependent classifiers $p_n^{(1)}, \dots, p_n^{(N)}$ trained on the same data $D_n = (X_1, Y_1), \dots, (X_n, Y_n)$. And suppose we have m additional data points $D'_m = (X'_1, Y'_1), \dots, (X'_m, Y'_m)$ to test the performance. Then, we may define

$$R'_n(p_n^{(j)}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{p_n^{(j)}(X'_i)\neq Y'_i}. \tag{48}$$

Now the classifiers do not depend on the test data, $R'_m(p_n^{(j)})$ will depend on D_n and D'_m, and $R(p_n^{(j)})$ will depend only on D_n.

$$|R'_m(p_n^{(j)} - R(p_n^{(j)})| \leqslant \max_{j=1,\dots,N} |R'_m(p_n^{(j)}) - R(p_n^{(j)})|. \tag{49}$$

If we condition now on D_n, then $p_n^{(j)}$ will be a fixed classifier not dependent on data, so we can bound it as

$$|R'_m(p_n^{(j)} - R(p_n^{(j)})| \leqslant \sqrt{\frac{\log \frac{2N}{\delta}}{2m}}, \tag{50}$$

with probability $\geqslant 1 - \delta$ respect to D'_m conditioned on D_n.

Vapnik–Chervonenkis (VC) theory

We will try to quantify overfitting by taking into account the structure of the class of classifiers. Given a class $\mathcal{C}$ of non-data-dependent classifiers, we want to understand

$$\max_{g \in \mathcal{C}} |R_n(g) - R(g)|. \tag{51}$$

We have seen that if $\mathcal{C}$ is finite, then with probability $\geqslant 1 - \delta$ the maximum is $\leqslant \sqrt{\frac{\log \frac{2N}{\delta}}{2n}}$. Observe that $R_n(g)$ counts for each g how many mistakes it takes, so it is a binomial random variable $R_n(g) \sim \frac{1}{n} Bin(n, R(g))$. So, we are looking for the deviation of binomials with their expected value.

Let's start simplifying the notation. Let $X_1, \ldots, X_n$ be iid taking values on Ω. For $A \subset \Omega$, we define $P(A) = P(X \in A)$ and $P_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \in A}$. And let $\mathcal{A}$ be a class of sets A. We are interested in

$$\max_{A \in \mathcal{A}} |P_n(A) - P(A)|. \tag{52}$$