Open guide for Machine Learning: Theory

Roger Garriga Calleja

February 1, 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

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

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

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

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

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 $P(S-\mathbb{E}S \geqslant t) \leqslant \min_{\lambda>0} \frac{\prod_{i=1}^{n} \mathbb{E}[e^{\lambda(X_i-\mathbb{E}X_i)}]}{e^{\lambda t}}$, now using Hoeffding's Lemma, $\min_{\lambda>0} \frac{\prod_{i=1}^{n} \mathbb{E}[e^{\lambda(X_i-\mathbb{E}X_i)}]}{e^{\lambda t}} \leqslant \min_{\lambda>0} \frac{e^{\lambda^2 n(b-a)}}{\lambda t} = \min_{\lambda>0} e^{\frac{\lambda^2 n(b-a)}{8} - \lambda t}$. 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 $v = \sum_{i=1}^{n} \mathbb{E}[X_i^2]$. 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 $\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, 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 $\{x_1, \ldots, x_l\}, \{x_{l+1}, \ldots, x_{2l}\}, \ldots \{x_{(K-1)l+1}, \ldots, x_{Kl}\}$, the means $\mu_1 = \frac{1}{l}\sum_{i=1}^{l} x_i, \mu_2 = \frac{1}{l}\sum_{i=l+1}^{2l} x_1, \ldots, \mu_K = \sum_{i=(K-1)l+1}^{Kl} x_i$ and the estimator $\hat{m}_n = \text{median}(\mu_1, \ldots, \mu_k)$.

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 $j = 1, \ldots, K$ (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 $\mu_1, \ldots, \mu_K$ 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 $p^* = \underset{p \in T}{\operatorname{argmin}} R(p)$. 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 $Med(Y) = \underset{p}{\operatorname{argmin}} \|Y - p\|^2$.

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

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