Open guide for Machine Learning: Theory

Roger Garriga Calleja

January 13, 2017

1 Important probability's inequalities

Markov's inequality

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

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

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 $\square$

Chebyshev's inequality

For any random variable with finite variance

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

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 $\square$

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,

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

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

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

Chernoff's bounds

For any $\lambda > 0$,

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

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 $\square$

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:

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

(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 $v = \sum_{i=1}^{n} \mathbb{E}[X_i^2]$. Then, $\forall t > 0$,

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

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

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

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 $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ 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 $\{x_1, \dots, x_l\}, \{x_{l+1}, \dots, x_{2l}\}, \dots \{x_{(K-1)l+1}, \dots, x_{Kl}\}$, the means $\mu_1 = \frac{1}{l}\sum_{i=1}^{l} x_i, \mu_2 = \frac{1}{l}\sum_{i=l+1}^{2l} x_1, \dots, \mu_K = \sum_{i=(K-1)l+1}^{Kl} x_i$ and the estimator $\hat{m}_n = \text{median}(\mu_1, \dots, \mu_k)$.

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

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

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:

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

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,

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

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.