Machine Learning Exercises: Set 1

Roger Garriga Calleja

January 27, 2017

Problem 1: Let X be a real-valued random variable with mean m, median M, and standard deviation. Prove that

$$|m - M| \leqslant \sqrt{2}\sigma.$$

Chebyshev's inequality states that $P(|X - \mathbb{E}X| \geqslant t) \leqslant \frac{\sigma^2}{t^2}$
First of all consider $|X - m|$ and apply Chevyshev with strict inequality and $t = \sqrt{2}\sigma$

$$P(|X - m| > \sqrt{2}\sigma) < \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}.$$

That means that $P(X \notin (m - \sqrt{2}\sigma, m + \sqrt{2}\sigma)) < \frac{1}{2}$. Taking the complementary we get that $P(X \in [m-\sqrt{2}\sigma, m+\sqrt{2}\sigma]) \geqslant \frac{1}{2}$, which means that the median will be in the region $M \in [m-\sqrt{2}\sigma, m+\sqrt{2}\sigma]$ $\Rightarrow |m - M| \leqslant \sqrt{2}\sigma$.

Problem 2: Write a program that compares the performance of the empirical mean and the median-of-means mean estimators. Test them on randomly generated samples drawn from different distributions, including heavy-tailed ones. Try different parameters of the median-of-means estimator and different sample sizes. Compare the estimators according to different measures, such as average deviation from the true mean, as well as worst case deviation (when the random sample is re-drawn many times). You may consider the Pareto family or Student's t-distribution (with different degrees of freedom) for heavy-tailed examples.

Problem 3: Let $X_1, \dots, X_n$ be i.i.d. *non-negative* random variables with mean $\mathbb{E}X_1 = m$ and second moment $\mathbb{E}X_1^2 = a^2$. Use the Chernoff bound to prove that, for all $t \in (0, m)$,

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} X_i < m - t\right\} \leqslant e^{-\frac{nt^2}{2a^2}}.$$

***Hint:* use the fact that for $x > 0$, $e^{-x} \leqslant 1 - x + \frac{x^2}{2}$.**

Working a bit the equation we get

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i < m - t\right) = P\left(t < m - \frac{1}{n}\sum_{i=1}^{n} X_i\right) = P\left(m - \frac{1}{n}\sum_{i=1}^{n} X_i > t\right) = P\left(\sum_{i=1}^{n} \mathbb{E}X - \sum_{i=1}^{n} X_i > nt\right).$$

Now we can apply Chernoff bound taking into account that X_i are independent, so

$$P\left(\sum_{i=1}^{n} \mathbb{E}X - \sum_{i=1}^{n} X_i > nt\right) \leqslant \frac{\mathbb{E}\prod_{i=1}^{n} e^{\lambda(\mathbb{E}X_i - X_i)}}{e^{\lambda tn}}.$$

Then, as we have $\mathbb{E}X_i = m \ \forall i$, we get

$$P\left(\sum_{i=1}^{n}\mathbb{E}X - \sum_{i=1}^{n} X_i > nt\right) \leqslant \frac{\mathbb{E}\prod_{i=1}^{n} e^{\lambda(\mathbb{E}X_i - X_i)}}{e^{\lambda tn}} = \frac{1}{e^{n\lambda t}}\prod_{i=0}^{n}\mathbb{E}e^{\lambda(m-X_i)} = \frac{e^{\lambda m}}{e^{n\lambda t}}\prod_{i=0}^{n}\mathbb{E}e^{-\lambda X_i}.$$

As $e^{-x} \leqslant 1 - x + \frac{x^2}{2}$ for $x > 0$ and our X_i are non-negative, we get

$$\frac{e^{\lambda m}}{e^{n\lambda t}}\prod_{i=0}^{n}\mathbb{E}e^{-\lambda X_i)} \leqslant e^{n\lambda(m-t)}\prod_{i=1}^{n}(1 - \lambda\mathbb{E}X_i + \frac{\lambda^2}{2}\mathbb{E}X_i^2).$$

Now, since $1 + x \leqslant e^x \ \forall x \geqslant 0$ (on $x = 0$ both are equal and e^x derivative is greater or equal to 1 for $x > 0$), we get

$$e^{n\lambda(m-t)}\prod_{i=1}^{n}(1 - \lambda\mathbb{E}X_i + \frac{\lambda^2}{2}\mathbb{E}X_i^2) \leqslant e^{n\lambda(m-t)}\prod_{i=1}^{n} e^{-\lambda m + \frac{\lambda^2}{2}a^2} = e^{n\lambda(m-t)}e^{n(-\lambda m + \frac{\lambda^2}{2}a^2)} = e^{-n\lambda t + n\frac{\lambda^2}{2}a^2}.$$

We minimize the function on λ to get the bound. $0 = (-n\lambda t + n\frac{\lambda^2}{2}a^2)' = -nt + n\lambda a^2 \Leftrightarrow \lambda = \frac{t}{a^2}$. That implies

$$e^{-n\lambda t + n\frac{\lambda^2}{2}a^2} \leqslant e^{-\frac{t^2}{2a^2}}.$$

So, $P\left(\frac{1}{n}\sum_{i=1}^{n} X_i < m - t\right) \leqslant e^{-\frac{nt^2}{2a^2}}$. Q.E.D.

Problem 4: Write a program that projects the n standard basis vectors in $\Re^n$ to a random 2-dimensional subspace. (You may do this simply by using a $2 \times n$ matrix whose entries are i.i.d. normals.) Center the point set appropriately and re-scale such that the empirical variance of the (say) first component equals 1. Plot the obtained point set. Now generate n independent standard normal vectors on the plane and compare the two plots. Do this for a wide range of values of n. What do you see?
Repeat the same exercise but now projecting the $2n$ vertices of the hypercube $\{1, 1\}^n$ instead of the standard basis vectors. (Naturally, you can only do this for small values of n, say up to $n \simeq 13$.) What do you see now?