Machine learning@Data science, 2019–Exercises.

Set 1. Due January 31, 2019

Problem 1 Write a program that compares the performance of three mean estimators: empirical mean, trimmed mean, and the median-of-means mean estimator. (The trimmed mean of n real numbers $x_1, \ldots, x_n$ is obtained by deleting the k smallest and k largest numbers—where $k < n/2$—and computes the empirical mean of the remaining $n - 2k$ numbers.)

To evaluate the performance of an estimator m_n, generate n i.i.d. random variables, compute m_n, and generate a large number of independent data points to estimate $\mathbf{P}\{|m_n - m| > \epsilon\}$ for a wide range of choices of ϵ and n. For the median-of-means estimator try various values of the block size and similarly for the trimmed-mean estimator.

Generate distributions for both light (such as Gaussian, Laplace) and heavy tailed distributions (such as the Pareto family or Student's t-distribution with different degrees of freedom).

Problem 2 Let X be a random vector uniformly distributed in the d-dimensional cube $[-1, 1]^d$ (i.e., the components of $X = (X_1 \ldots, X_n)$ are independent, uniformly distributed in the interval $[-1, 1]$). What can you say about the distribution of $\|X\|^2$? Determine the mean, the variance, and establish concentration inequalities.

If X' is another independent vector drawn from the same distribution, what is the "typical" order of magnitude (as a function of d) of the cosine of the angle between X and X'? Recall that the cosine of the angle between two vectors u and v is

$$\frac{u^T v}{\|u\| \cdot \|v\|} \ .$$

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

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

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

Problem 4 Write a program that projects the n standard basis vectors in $\mathbb{R}^n$ 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?