Machine learning@Data science, 2021–Exercises.

Set 1. Due January 29, 2021

Problem 1 The median-of-means estimator is not permutation-invariant in the sense that if the indices of the data $X_1, \ldots, X_n$ are permuted to $X_{\sigma(1)}, \ldots, X_{\sigma(n)}$ for a permutation $\sigma = (\sigma(1) \ldots, \sigma(n))$, then the value of the estimator may change. Denote the median-of-means estimator based on $X_{\sigma(1)}, \ldots, X_{\sigma(n)}$ by $m_n(\sigma)$.

In order to define a permutation-invariant version, we may simply take the average over all permutations:

$$\overline{m}_n = \frac{1}{n!} \sum_{\sigma} m_n(\sigma) \ ,$$

where the sum is taken over all $n!$ permutations. Unfortunately, computing a sum of $n!$ terms is impossible even for small values of n. One may try to approximate the average by Monte-Carlo sampling. More precisely, let $\sigma_1, \ldots, \sigma_N$ be independent random permutations and define

$$\widehat{m}_{n,N} = \frac{1}{N} \sum_{j=1}^{N} m_n(\sigma_j) \ .$$

How large does N have to be in order to get a good approximation of $\overline{m}_n$? Use concentration inequalities to quantify your answer.

Problem 2 Write a program that compares the performance of three mean estimators: empirical mean, median-of-means mean estimator, and the permutation-invariant median-of-means mean estimator described in the previous exercise (using Monte-Carlo approximation).

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 for the permutation-invariant version examine the effect of the parameter N.

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 3 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 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 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?