EECS 189 Introduction to Machine Learning
Fall 2020 DIS2

This discussion was released **Friday, September 11**.

1 Overview of test sets, validation, and cross-validation

In this part, we discuss several issues having to do with test sets and the notions of validation and cross-validation. Open this notebook in datahub and discuss the questions it contains.

The following is Problem 4.d from HW2. This is a standalone problem, i.e., it does not depend on results from 4.a–4.c.

2 Outlier Removal via OMP (Part 2)

(a) From the law of large numbers, we have seen that with a large number of samples, the sample mean converges to the population mean or expected value. More rigorously, the weak law of large numbers states the following: For any positive number ε,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) = 0$$

where μ is the expectation, and $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. Here, we would like to make a similar statement for sample median and population median. Given a sequence of n random variables i.i.d. drawn from the same distribution, $\{X_1, X_2, ..., X_n\}$, let's denote the population median as $med(X)$ and the sample median as $\tilde{X}_n$. We want to make no other assumption on the distribution of X_i's. The goal is to give a proof of the following statement:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\tilde{X}_n - med(X)\right| > \epsilon\right) = 0$$

But to make the proof easier to follow and to understand things in terms of their natural dependencies, we will modify the above statement to involve *quantiles* of X. For every $\varepsilon > 0$ for which the $(\frac{1}{2} - \varepsilon)$ quantile is different from the median and the $(\frac{1}{2} + \varepsilon)$ quantile is also different from the median, we have:

$$\lim_{n\to\infty} \mathbb{P}\left(\tilde{X}_n < (1/2 - \varepsilon)\text{-quantile or } \tilde{X}_n > (1/2 + \varepsilon)\text{-quantile}\right) = 0.$$

Here, (for simplicity) a p-quantile of a random variable is a value x for which the CDF $\mathbb{P}(X \le x) = p$. [To be precise, a p-quantile is an x for which $\mathbb{P}(X < x) \le p$ and $\mathbb{P}(X \le x) \ge p$. This allows the distribution of X to have atoms in it and for quantiles to still be defined in a reasonable manner.] Notice that by choosing an appropriate value of ε, we can recover the desired ϵ, and hence, the two statements are equivalent. *(First hint: Consider a Bernoulli random variable: $Y_i = \mathbb{1}\{X_i > (1/2 + \varepsilon)\text{-quantile}\}$) (Second hint: Think about a relevant Chernoff bound* *and use it. You don't have to use a Chernoff bound to prove it, but it helps in understanding the speed of this convergence.)*

Contributors:

- Anant Sahai
- Chawin Sitawarin
- Inigo Incer