ECE 6980
An Algorithmic and Information-Theoretic Toolbox for Massive Data

Instructor: Jayadev Acharya — Lecture #7
Scribe: JA — 15th September, 2016

Please send errors to acharya@cornell.edu

1 Robust Estimation

Until now, we assumed the underlying distribution P belongs to a known class $\mathcal{P}$, and we observe samples $X_1^n \sim P$. However, the assumption that $P \in \mathcal{P}$ might not be true, or it is possible that some of the entries are corrupted. Learning distributions under various possible corruptions is called as *robust statistics.*

1.1 Huber's contamination model

The most widely used model is Huber's contamination model. P_u is an underlying distribution, and P_e is an error distribution. Suppose we observe n samples X_1^n where each X_i is distributed:

$$X_i \sim (1-\varepsilon)P_u + \varepsilon P_e.$$

In other words, with probability ε the sample is not generated from P_u, but from P_e.

1.2 Adversarial contamination

A slightly stronger model is the following. First generate X_1^n from the underlying distribution P_u. An adversary can select *any* $X_{j_1}, \ldots X_{j_{n\varepsilon}}$ and replaces them with *anything they like.*

The goal is to estimate the underlying distribution $P_u \in \mathcal{P}$ up to a total variation distance $C\varepsilon$ for some absolute constant C.

Let us consider the following concrete question.

Question. Suppose $P_u = N(\mu, 1)$. The goal is to output a distribution $\hat{P}$ such that $d_{TV}(P_u, \hat{P}) \leq C\varepsilon$, for some constant ε. We want to understand:

- How many samples are required?
- What is the time complexity?

In Assignment 1 Problem 3, when we substitute $\sigma = 1$, we showed that if $\mu - hat\mu < c_1\varepsilon$ (for some constant ε), there is a constant C_1 (depending on c_1) such that

$$d_{TV}((,N)(\mu,1), N(\hat{\mu},1)) \leq C_1\varepsilon.$$

Using this we showed that $N((\sum X_i)/n, 1)$ can estimate $N(\mu, 1)$ to total variation $O(\varepsilon)$ using $O(1/\varepsilon^2)$ samples. Please attempt that problem (first two parts) to get a sense of the sample complexity of C/ε^2.

Consider another approach. Let X_{med} denote the median of X_1^n.

We studied the CDF of a distribution last time in class. In particular we stated that:

Claim 1. *Suppose P be any distribution on $\mathbb{R}$, and P_n be the empirical distribution upon observing n samples. Let $F_P(x) = P(X \le x)$ denote the CDF of P. Then for any x,*

$$\mathbb{E}[|F_{P_n}(x) - F_P(x)|] \le \sqrt{1/n}).$$

This result follows from the fact that $F_{P_n}(x)$ is the number of symbols that are at most x, which is a Binomial distribution with parameters n and $F_P(x)$. The remaining part follows from the computations we did for variances of Binomials (refer to previous notes on estimating Bernoulli random variables).

We first state a simple result for Gaussian distributions, which is left as an exercise (along the lines done in class).

Claim 2. *Let $P = N(\mu, 1)$, and $\varepsilon > 0$ a small constant (say < 0.1). Then there are universal constants C_l, C_u, such that*

$$1/2 - C_l\varepsilon < P(X \le \mu - \varepsilon) < \frac{1}{2} - C_u\varepsilon.$$

Using this, we prove the following claim:

Claim 3. *There is a constant C such that given C/ε^2 samples, with probability at least 0.9 the median lies between $\mu - \varepsilon$, and $\mu + \varepsilon$.*

Proof. By Claim 1, we can choose C to be such that for $x = \mu - \varepsilon$ and $x = \mu + \varepsilon$, w.p. at least 0.9, (using Markov Inequality),

$$|F_{P_n}(\mu \pm \varepsilon) - F_P(\mu \pm \varepsilon)| < \frac{\varepsilon C_u}{10}.$$

Therefore, the number of points less than $\mu - \varepsilon$ is at most $n(1/2 - C_u\varepsilon) + C_u\varepsilon n/10 < n/2$, and the number of points larger than $\mu + \varepsilon$ is more than $n/2$. Therefore the median lies in the interval. □

This proves that the median is a good substitute for the sample mean. We can show that even with corrupting ε fraction of the symbols, the median does not change by more then $O(\varepsilon)$. In particular, one can show that there exist constants C_1 such that

$$1/2 - 4\varepsilon < P(X \le \mu - C_l\varepsilon).$$

By the same argument as the claim, we can show that with for $n > C/\varepsilon^2$ the number of samples at most $\mu - C_l\varepsilon$ is at most $n/2 - 3n\varepsilon$, and the number of elements at most $\mu + \varepsilon$ is at least $n/2 + 3n\varepsilon$. Therefore, there are at least $6n\varepsilon$ elements are within $C_1\varepsilon$ distance of μ. Therefore, when we change $n\varepsilon$ fraction of the elements, the median does not leave this interval, and we are done.

1.3 Higher dimensions

Consider the problem of learning Gaussians in d dimensions, with unit covariance matrix. Namely, the $P_u = N(\overline{\mu}, \mathbb{I}_d)$. Again, a fraction ε of the samples are corrupted. We noticed in the assignment that if there are no corruptions $O(d/\varepsilon^2)$ samples are sufficient to learn such a P_u.

If we allow some corruptions, one can construct large collection of distributions by picking each coordinate to be any of the X_{ij}'s along the lines of what we did for learning mixtures of Gaussians. If we do this, then a covering argument will give "sample efficient", but exponential time algorithms. We do not want that. We briefly mentioned about two papers that avoid this exponential time.