Lecture 1: Introduction to robustness

October 2, 2019

1 Robustness in machine learning

Robustness in the context of machine learning can mean a million things to a million people. In this course, we wish to try to rigorously study the role of corruption in machine learning. To do so, it will be helpful to consider the three following questions:

1.1 What types of corruptions can occur?

Indeed, corruptions can occur at essentially every part of the machine learning pipeline, and of course, multiple types of corruption can occur simultaneously. In this course, we will consider corruptions at training time, test time, and in the modeling of the problem itself.

Beyond this, we also need to specify what sorts of corruptions we will consider. Classically, we often model corruption as white noise: i.e. rather than observing x, we observe $x + \eta$, where η is a mean zero random variable. While this is a useful heuristic for a number of settings, in recent years it has become apparent that in many situations it is important to study *worst case* noise. This is the type of corruption that we will mostly focus on in this class.

Even so, there is more to specify. We will consider both *gross corruption*, in which a small (or even large) fraction of the data points are completely corrupted, as well as *constrained corruptions*, in which potentially all of the data points have been altered, but in some restricted way. And of course there are many ways to defined what constraints are allowed. In general, there is no "one-size-fits-all" notion of corruption, and the types of corruptions that one may wish to consider will depend on the problem at hand.

1.2 What do we know about the uncorrupted data?

It is a general rule of thumb that unless there is some structure amongst the original, uncorrupted data, then we cannot really say anything about the robust learning task. This is simply because then we cannot even distinguish whether or not the data we see is the original data, or whether it is the corrupted data. More generally, the amount of structure we assume will dictate the statistical limits of what is possible to recover in the presence of noise, and in general, the more structure we assume, the more information can be recovered. Of course, there is a tradeoff, as the more restrictive the generative model is, the fewer situations it holds in. It is again a very domain-specific question of what sorts of assumptions we can make.

1.3 What information do we want to recover?

We now have a well-defined generative model: we know how the uncorrupted data is created, and how it may be corrupted. To finish defining the problem, it suffices to understand what information we wish to recover. As mentioned previously, the presence of worst-case corruptions typically means that we cannot hope to recover all information. Properly defining which statistics we wish to recover, and what notion of closeness we wish to use, will be crucial to getting interesting results. As we shall see, this will often crucially depend on both the assumptions on the corrupted and uncorrupted data we are making.

2 Robust statistics

The first topic we will cover in this class is *robust statistics*. Broadly defined, this is the study of learning in the presence of corrupted training data. We will cover a variety of concrete settings in which one may wish to do this, starting from the most basic, to more real-world settings near the end of this unit. The basic question, which dates back to work of Huber and Tukey and others in the 60's and 70's [1, 2, 3, 4], asks the following:

Given samples from a distribution D, where an ε-fraction have been corrupted, when can we recover statistics (e.g. mean, covariance) of D?

2.1 Robust mean estimation

Specifically in this class (and the next couple of classes), we will consider *robust mean estimation*, arguably the most simple of these tasks. Here, the goal is simply (as the name suggests) to recover the mean of the distribution, given corruptions. Let's specify this problem more carefully, by answering the three questions posed above. We will go slightly out of order:

What types of corruptions can occur? In robust statistics, the types of corruptions considered are gross corruptions. Namely, we assume that an ε-fraction of our data can be *arbitrarily* corrupted. Even here we can have differences. In this class, for simplicity, we will consider *additive* corruptions. We assume that the adversary has simply added an ε-fraction of corrupted data, but cannot remove them. Later on, we will also consider adversaries which can remove good data, as well as more statistical notions of worst case noise.

What information do we want to recover? As this is robust mean estimation, we wish to recover the mean of the distribution D. In this class, we will focus on univariate distributions, so the measure of distance is simply distance to the true mean μ along the real line. In future classes, we will consider multivariate settings, where the measure of distance will change, depending on the problem setting.

What do we know about the uncorrupted data? We will consider two representative settings. The first, and perhaps most constrained, is when we make the relatively strong assumption that the distribution D is Gaussian, with known variance. As we shall see, the fact that we know the variance is not so much an issue. More so the problem is that Gaussians are very structured distributions, with strong concentration properties (amongst other nice things). To attempt to alleviate this, we will also consider the much weaker assumption, namely, that D only has bounded variance. As we shall see, while this may be a more realistic assumption, the guarantees we can achieve in this setting are much weaker than in the Gaussian setting.

2.2 Problem statements

We have now two formal mathematical problems to consider:

Problem 2.1 (Robustly learning the mean of a Gaussian under additive corruptions)**.** *Let $\mu \in \mathbb{R}$, and let $\varepsilon < 1/2$. Let S be a set of n samples so that $S = S_{\text{good}} \cup S_{\text{bad}}$, where S_{good} is a set of i.i.d. samples from $\mathcal{N}(\mu, \sigma^2)$, and S_{bad} satisfies $|S_{\text{bad}}| < \varepsilon n$. Given S, ε, σ, output $\widehat{\mu}$ minimizing $|\mu - \widehat{\mu}|$.*

Problem 2.2 (Robustly learning the mean of distribution with bounded second moments under additive corruptions)**.** *Let $\mu \in \mathbb{R}$, and let $\varepsilon < 1/2$. Let S be a set of n samples so that $S = S_{\text{good}} \cup S_{\text{bad}}$, where S_{good} is a set of i.i.d. samples from D, and S_{bad} satisfies $|S_{\text{bad}}| < \varepsilon n$, where D has variance at most σ^2. Given S, ε, σ, output $\widehat{\mu}$ minimizing $|\mu - \widehat{\mu}|$.*

When $\varepsilon = 0$, i.e., when there are no corruptions, the empirical mean is solves these problems well. However, when there is any corruption, the empirical mean fails horribly (why?). The rest of this lecture will be dedicated to the analysis of two algorithms, one for each problem, which do tolerate corruptions, namely, the

median and the *truncated mean*. Both of these basic algorithms will demonstrate some high level principles which will be useful as we move towards more complicated versions of this problem.

2.3 Robust mean estimation for Gaussians via medians

An algorithm which will work for Problem 2.1 is the median of our dataset, which we will denote $\mathrm{med}(S)$. Recall that Φ is the cdf of the standard normal Gaussian:

$$\Phi(t) = \Pr_{X \sim \mathcal{N}(0,1)}[X \le t] \; .$$

We will show the following guarantee for the median:

Theorem 2.3. *Let $\mu, \varepsilon, \sigma, S$ be as in Problem 2.1, let $n = |S|$, and let t be so that $t \ge \Phi^{-1}(1/2 + \varepsilon)$. Then:*

$$\Pr\left[|\mathrm{med}(S) - \mu| > t\sigma\right] \le 2\exp(-2n(\Phi(t) - 1/2 - \varepsilon)^2) \; . \tag{1}$$

Before we prove this theorem, we make two observations. First, observe that if $\varepsilon \ge 1/2$, the condition on t is unsatisfiable. This makes sense: when there are more bad points than good points, we cannot hope to get a single consistent answer. However, for every $\varepsilon < 1/2$, we get some finite value of t at which the bound becomes nontrivial. This is related to the *breakdown point* of the estimator, as we will define later.

Secondly, we observe that when ε, t are relatively small, then we have

$$\Phi^{-1}(1/2 + \varepsilon) = \sqrt{2\pi}\varepsilon + O(\varepsilon^2) = (1 + o(1))\sqrt{2\pi}\varepsilon$$
$$\Phi(t) = \frac{1}{2} + \frac{t}{\sqrt{2\pi}} + O(t^2) = \frac{1}{2} + (1 + o(1))\frac{t}{\sqrt{2\pi}} \; .$$

Thus, we have the following:

Corollary 2.4. *Let $\mu, \varepsilon, \sigma, S$ be as in Problem 2.1, let $n = |S|$, and let $c > 0$ be a sufficiently small universal constant. Assume that $\varepsilon < c$. Then, for all $b > 0$ sufficiently small, we have*

$$\Pr\left[|\mathrm{med}(S) - \mu| > (1 + b)\sqrt{2\pi}\varepsilon\sigma\right] \le 2\exp(-\Omega(bn\varepsilon^2)) \; .$$

In particular, this says that with high probability (one can also prove expectation bounds), then given $n = \Omega(1/\varepsilon^2)$ samples, we can learn the mean robustly to error $O(\varepsilon)$. As we shall see in the next lecture, this is optimal: no algorithm, given any number of corrupted samples, can hope to do better than $\Omega(\varepsilon)$ error.

Proof of Theorem 2.3. By scaling, we may assume without loss of generality that $\sigma = 1$. We will show that

$$\Pr\left[\mathrm{med}(S) - \mu > t\sigma\right] \le \exp(-2n(\Phi(t) - 1/2 - \varepsilon)^2) \; ; \tag{2}$$

the argument for bounding the lower tail is symmetric, and we just lose a factor of two by a union bound.

Recall that we may write S as the disjoint union of S_{good} and S_{bad}, where $|S_{\mathrm{good}}| \ge (1 - \varepsilon)n$, and the points in S_{good} are drawn i.i.d. from $\mathcal{N}(\mu, 1)$. Observe that the median of S is at most the $(1/2 + \varepsilon)$-quantile of S_{good}, since S_{bad} contains only an ε-fraction of points. Hence it suffices to show that the $(1/2 + \varepsilon)$-quantile of S_{good} is not too large.

For each $i \in S_{\mathrm{good}}$, let Y_i be the $\{0, 1\}$-valued random variable which is 1 if $X_i - \mu > t$, and 0 otherwise. Notice that the Y_i are i.i.d. Bernoulli random variables, and

$$\mathbb{E}[Y_i] = \Phi(-t) = 1 - \Phi(t) \; .$$

Moreover, the $(1/2 + \varepsilon)$-quantile of S_{good} exceeds $\mu + t$ if and only if $\frac{1}{n}\sum_{i \in S_{\mathrm{good}}} Y_i \ge 1/2 - \varepsilon$. By a Chernoff bound, we have that for all $s > 0$,

$$\Pr\left[\frac{1}{n}\sum_{i \in S_{\mathrm{good}}} Y_i > 1 - \Phi(t) + s\right] \le \exp(-2ns^2) \; .$$

Let $s = \Phi(t) - 1/2 - \varepsilon$. Observe that by assumption, we have that $s > 0$, and so we may apply the above bound with this choice of s. By doing so, we obtain the desired claim. □

2.4 Robust mean estimation with bounded moments via truncated mean

We now consider the second problem: namely, here we will assume that the good distribution D has variance at most σ^2. This is a substantially weaker assumption than Gaussianity; consequently the error we get will also be much weaker. The algorithm we will use here is quite intuitive: remove the top 2ε-quantile of samples, as well as the bottom 2ε-quantile of samples, and then take the mean of the rest. Given a set of samples $S = \{X_1, \dots, X_n\}$, we let $\mathrm{tmean}_\varepsilon(S)$ denote the trimmed mean of this set of samples. The remainder of this section is dedicated to the proof of the following theorem:

Theorem 2.5. *Let $D, \mu, \varepsilon, \sigma, S$ be as in Problem 2.2, and let $\varepsilon < 1/4$. Then, with probability $9/10$, we have*

$$|\mathrm{tmean}_\varepsilon(S) - \mu| \lesssim \sigma\sqrt{\varepsilon} + \sqrt{\frac{\sigma^2}{n}} \,.$$

The crucial fact that we will use to prove this theorem will be the following concentration inequalities about random variables with bounded second moment.

Fact 2.6. *Let D be a distribution with variance bounded by σ^2 and mean μ. Then:*

- *(Chebyshev's inequality) If X is distributed as D, then $\Pr[|X - \mu| > t] \le \sigma^2/t^2$, for all $t > 0$.*
- *For any event E, we have*

$$\left|\mathop{\mathbb{E}}_{D}[(X-\mu)\mathbb{I}_E]\right| \le \sigma \Pr_D[E]^{1/2} \,.$$

Proof. The first statement follows from Markov's inequality applied to the random variable $(X-\mu)^2$. We now prove the second statement. We have:

$$\left|\mathop{\mathbb{E}}_{D}[(X-\mu)\mathbb{I}_E]\right| \le \left(\mathop{\mathbb{E}}_{D}\left[(X-\mu)\right]^2 \mathop{\mathbb{E}}_{D}[\mathbb{I}_E]\right)^{1/2} = \sigma \Pr_D[E]^{1/2} \,.$$

Here the main inequality follows from the fact that $\mathbb{E}[f]^2 \le \mathbb{E}[f^2]$. □

Proof sketch of Theorem 2.5. Notice that no matter what the adversary does, we always remove the top ε-quantile of points and bottom ε-quantile of points from S_{good}. If we let E be the event $E = \{X \in [\mu - C\sigma/\sqrt{\varepsilon}, \mu + C\sigma/\sqrt{\varepsilon}]\}$, then observe that $\Pr_D[X \notin E] \le \varepsilon/2$, by Chebyshev's inequality, for some C sufficiently large. By a concentration argument similar to what we did for median above, with high probability (say $\ge 99/100$), the largest remaining point in S_{good} will be at most $\mu + C\sigma/\sqrt{\varepsilon}$ and the smallest remaining point will be at least $\mu - C\sigma/\sqrt{\varepsilon}$. In particular, this means that any point (including those from S_{bad}) that survives the truncation is at most $C\sigma/\sqrt{\varepsilon}$ away from μ.

We now observe that

$$\left|\mathop{\mathbb{E}}_{D}[X|E] - \mu\right| = \left|\mathop{\mathbb{E}}_{D}[(X-\mu)|E]\right| = \frac{1}{\Pr[E]}\left|\mathbb{E}\left[(X-\mu)\mathbb{I}_E\right]\right| = \frac{1}{\Pr[E]}\left|\mathbb{E}[(X-\mu)\mathbb{I}_{E^c}]\right| \lesssim \sigma\sqrt{\varepsilon} \,,$$

by Fact 2.6. Moreover, $\mathbb{E}_D[(X-\mu)^2|E] \le \frac{\sigma^2}{\Pr[E]} \lesssim \sigma^2$. Thus the random variable X conditioned on the event E has bounded second moment and its mean differs from μ by at most $\sigma\sqrt{\varepsilon}$. Combining these two facts and Chebyshev's inequality, we have

$$\Pr\left[\left|\frac{1}{|S_{\mathrm{good}} \cap E|}\sum_{i \in S_{\mathrm{good}} \cap E} X_i - \mu\right| \gtrsim \sigma\sqrt{\varepsilon} + \sqrt{\sigma^2/n}\right] < 1/100 \,.$$

Condition on this event happening as well. Now we are almost done. Let $T \subseteq S$ be the set of points that survive the truncation. We can decompose T into two sets: $T \cap S_{\text{good}} \cap E$ and $T \cap S_{\text{bad}}$. We further have

$$\begin{aligned}\left|\sum_{i \in T \cap S_{\text{good}} \cap E} (X_i - \mu)\right| &\le \left|\sum_{i \in S_{\text{good}} \cap E} (X_i - \mu)\right| + \left|\sum_{i \in S_{\text{good}} \cap E \cap T^c} (X_i - \mu)\right| \\ &\lesssim |S_{\text{good}} \cap E| \cdot (\sigma\sqrt{\varepsilon} + \sqrt{\sigma^2/n}) + \sigma\sqrt{\varepsilon} n \ ,\end{aligned}$$

where the inequality follows because of the event we are conditioning on, and since we throw out at most $2\varepsilon n$ points, and each point can contribute at most $\sigma/\sqrt{\varepsilon}$ (up to constants). Similarly we have

$$\left|\sum_{i \in T \cap S_{\text{bad}}} (X_i - \mu)\right| \lesssim \sigma\sqrt{\varepsilon} n \ .$$

Putting it all together with triangle inequalities and dividing through by $|T|$ yields the desired conclusion. □

Observe this proof can also work for other concentration assumptions. For instance, if we have bounded k-th moments, we get error that scales more like $\sigma\varepsilon^{1-1/k}$.

References

[1] Frank J Anscombe. Rejection of outliers. *Technometrics*, 2(2):123–146, 1960.

[2] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

[3] John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531, 1975.

[4] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.