Lecture 1: Upper bound on learning unstructured distribution

Daniel Kane
Scribe: Sankeerth Rao

April 3, 2017

Abstract

This lecture introduces the basic set up of distribution learning and proves an upper bound on the number of samples required for learning an unstructured distribution.

1 Setup

Say there is an unknown probability distribution p (perhaps known to satisfy extra properties). We take independent samples from p and would like to determine some information about p.

The main parameters of this algorithm that we need to keep track of are:

- **How many samples ?** We want almost information theoretically optimal (within constant factors)
- **How efficient is the algorithm ?** Ideally we want near linear in the number of samples, but we also accept polynomial time algorithms.
- **What is the probability of failure ?** We require usually require only a $\frac{2}{3}$ probability of success, but this doesn't matter very much. We can usually amplify the probability of success to $1-\delta$ with $O(\log(1/\delta))$ independent repetitions of the same algorithm.

Let us begin with the following example:

2 Learning Unstructured distribution

Let p be an arbitrary distribution on $[n] = \{1, 2, \ldots, n\}$.

Objective: Learn p Note that this cannot be done exactly because there are infinitely many such distributions but we are only given access to a finite number of samples. So we revise ou objective as follows.

Revise: Return another distribution q such that

$$d_{TV}(p,q) = \frac{1}{2}|p-q|_1 = \frac{1}{2}\sum_{i=1}^{n}|p_i - q_i| < \epsilon$$

Intuition: A nice way to think of Total Variation distance is the by the following coupling inequality

Lemma 2.1. *Let μ, ν be two probability measures. For any rvs X, Y whose marginals are μ, ν we have*

$$||\mu - \nu||_{TV} \leq \Pr[X \neq Y]$$

In fact X, Y can be constructed so that this is an equality.

3 Algorithm:

The algorithm is very simple. Take N independent samples and we take the empirical distribution.

$$q_i = \frac{\text{No of samples in the ith bin}}{N}$$

4 Analysis:

Let X_i denote the number of samples from bin i. Then $q_i = \frac{X_i}{N}$.

The total variation distance is $d_{TV}(p,q) = \frac{1}{2}\sum_{i=1}^{n}|p_i - \frac{X_i}{N}|$.

Note that $X_i \sim Bin(p_i, N)$ is a Bernoulli random variable. Thus,

$$\mathbb{E}[X_i] = p_i N, \quad Var(X_i) = Np_i(1-p_i) < p_i N.$$

Thus, $\mathbb{E}[p_i - \frac{X_i}{N}] = 0$ and

$$\mathbb{E}\Big|p_i - \frac{X_i}{N}\Big|^2 = Var\Big(p_i - \frac{X_i}{N}\Big) \leq \frac{p_i}{N}.$$

Now using linearity of expectation we have,

$$\mathbb{E}\Big[\sum_i \Big|p_i - \frac{X_i}{N}\Big|^2\Big] \leq \frac{\sum_i p_i}{N} = \frac{1}{N}$$

This bounds $\mathbb{E}[||p-q||_2]$ but since we use Cauchy Schwarz + Jensen to get a bound on $\mathbb{E}[d_T V(p,q)]$.

$$\sum_i \mathbb{E}\left[\Big|p_i - \frac{X_i}{N}\Big|\right] \cdot 1 \leq \sqrt{\sum_i \left(\mathbb{E}\left[\Big|p_i - \frac{X_i}{N}\Big|\right]^2\right) \cdot \sum_{i=1}^{n} 1} \leq \sqrt{\sum_i \mathbb{E}\left[\Big|p_i - \frac{X_i}{N}\Big|^2\right] \cdot \sum_{i=1}^{n} 1} \leq \sqrt{\frac{n}{N}}.$$

Thus, we have $\mathbb{E}[d_T V(p,q)] \leq \sqrt{\frac{n}{N}}$. Say if we choose N so that $\sqrt{\frac{n}{N}} \leq \epsilon/3$ then using Markov inequality we have $d_T V(p,q) < \epsilon$ with probability at least 2/3. Thus,

$$N = O(\frac{n}{\epsilon^2})$$

is sufficient.

This proves the upper bound on the number of samples.

5 Lower bound

We need to prove that any algorithm that with prob 2/3 returns an $\epsilon-$approximation of p uses $>> \frac{n}{\epsilon^2}$ samples. We use information theory to prove this.

We use the adversary method. Let the adversary have a distribution $\mathcal{D}$ over all possible p. The algorithm gets N samples from a random $p \in \mathcal{D}$. We choose $\mathcal{D}$ wisely so that there is not enough information for the Algorithm to give the correct answer consistently. We prove this in the next Lecture.