1 Overview

Previously we have covered learning and testing on a unstructured discrete distribution. Notice that it is usually not possible to learn or distinguish unstructured continuous distribution with linear number of samples. For example it is impossible to distinguish the following two distributions with $n$ samples:

- $U_{[0, 1]}$ uniform distribution on $[0, 1]$.
- $U_S$ where $S$ is some random subset of $n^2$ points on $[0, 1]$.

Today we will cover some general techniques to learn structured continuous distributions. This method is usually not the most efficient algorithm, but gives information theoretical optimal sample complexity in a large fraction of time.

1.1 Formalize

The question can be formalized as following:

- Setup: Define $\mathcal{C}$ as a class of probability distribution, which might be a mixture of Gaussian, piecewise-linear, or other distributions.
- Goal: Assume $p \in \mathcal{C}$, distinguish $p$ with finite sample complexity.

2 Simple Case

2.1 Problem

Assume we have an unknown distribution $p$ either equals to $q_1$ or equals to $q_2$. We want to learn about $p$ in finite sample complexity.

2.2 Intuition

It roughly takes $\frac{1}{d_{tv(q_1, q_2)}}$ to learn distribution $p$. We need more samples if $q_1$ and $q_2$ are close to each other in terms of total variation distance.

Define $\epsilon = d_{tv(q_1, q_2)}$, and this notation will be used in following parts of this note. We have the following properties:
• Learning of \( p \) requires roughly \( \Omega\left(\frac{1}{\epsilon}\right) \) samples.

Notice that \( q_1 = q_2 \) with probability \( 1 - \epsilon \) because their total variation distance equals to \( \epsilon \). If we take samples from \( p = q_1 \), we need at least one sample from \( q_2 \) that differs from \( q_1 \) to distinguish, which has a \( \Omega(1/\epsilon) \) chance to happen. Note that this bound is not tight.

• \( O(1/\epsilon^2) \) is always sufficient to learn \( p \).

By definition of total variation distance, we have:

\[
d_{tv(q_1, q_2)} = \sup_A |q_1(A) - q_2(A)|
\]

Then with \( O(1/\epsilon^2) \) we are able to estimate \( P(A) \) to error of \( \epsilon/2 \), so it is enough to distinguish between \( q_1 \) and \( q_2 \) with are \( \epsilon \)-far from each other.

### 3 Generalization

After we had a taste of the first simple case, now it’s time to generalize the problem in a natural way.

#### 3.1 Problem

Assume we have \( q_1, q_2, \ldots, q_n \) distributions, and \( p \) is an unknown distribution that \( \epsilon \)-far from at least one of the \( n \) \( q \) distributions. In other words, \( \exists i \) s.t. \( d_{tv(p, q_i)} < \epsilon \).

#### 3.2 Tournament method

##### 3.2.1 Intuition

For each \( i \) and \( j \), define

\[
a_{ij} = d_{tv(q_i, q_j)} = |q_i(A_{ij}) - q_j(A_{ij})|
\]

where

\[
A_{ij} = \{x : q_i(x) \geq q_j(x)\}
\]

We estimate all of the \( P(A_{ij}) \) to error \( \delta \) with \( O\left(\frac{\lg(n)}{\delta^2}\right) \) samples to make them simultaneously correct pair-wise. Then we could union bound \( p \) with the estimations we calculated.

##### 3.2.2 Formalize

1. If \( d_{tv(p, q_i)} < \epsilon \):

\[
|\hat{p}(A_{ij}) - q_i(A_{ij})| \leq |p(A_{ij}) - \hat{p}(A_{ij})| + |p(A_{ij}) - q_i(A_{ij})| \\
\leq \epsilon + \delta
\]
where \( \hat{p} \) is the empirical distribution we estimated from samples. On the right hand side, the first term \( |p(A_{ij}) - \hat{p}(A_{ij})| < \delta \) assuming we estimated with enough sample, and the second term \( |p(A_{ij}) - q_i(A_{ij})| < \epsilon \) by assumption.

2. If \( d_{tv}(p,q_j) > 3\epsilon + 2\delta \):

\[
|\hat{p}(A_{ij}) - q_j(A_{ij})| \geq |p(A_{ij}) - q_j(A_{ij})| - \delta \\
\geq |q_j(A_{ij}) - q_i(A_{ij})| - |p(A_{ij}) - q_i(A_{ij})| - \delta \\
\geq d_{tv}(p,q_j) - d_{tv}(p,q_i) - \epsilon - \delta \\
\geq \epsilon + \delta
\]

During the second step, the first term \( |q_j(A_{ij}) - q_i(A_{ij})| = d_{tv}(q_i,q_j) \), and the second term \( |p(A_{ij}) - q_i(A_{ij})| \geq \epsilon \). We are relating \( q_i \) and \( q_j \) with \( p \) because \( p \) is close to one and far from the other.

During the third step, notice that \( d_{tv}(p,q_j) \leq 3\epsilon + 2\delta \) by assumption, and \( d_{tv}(p,q_i) < \epsilon \).

We call \( i \) good if \( |\hat{p}(A_{ij}) - q_i(A_{ij})| < \epsilon + \delta \), and \( i \) is not good if \( |\hat{p}(A_{ij}) - q_i(A_{ij})| > 3\epsilon + 2\delta \). If we take \( p = q_i \) with some good \( i \), then \( d_{tv}(p,q) < 3\epsilon + 2\delta \).

### 3.3 Conclusion

If we have such settings, we are able to find that specific hypothesis of \( p \) which is \( \epsilon \)-far from \( p \). If we are required to return a proper Hypothesis, then the 3-factor of \( \epsilon \) is required. While if we can return a mix of \( q \)'s as the hypothesis, then it is possible to achieve 2-factor \( \epsilon \)

The algorithm is very useful because it is a general technique to deal with structured distributions. Although the algorithm is not very efficient, it at least gives you a way to learn about these structured continuous distributions.

### 4 Covering

For any \( \mathcal{C} \) defined before, we are able to apply tournament method to learn distribution \( p \in \mathcal{C} \) if we are able to cover \( \mathcal{C} \) with finite \( \epsilon \)-balls, which usually has a logarithmic dependence on \( n \).

#### 4.1 Example 1

\[ \mathcal{C} = \{ N(\mu,\sigma) : \mu \in [-1,1], \sigma \in \left[ \frac{1}{10},10 \right] \} \]

\(^1\)I am not entirely sure the 3-factor and 2-factor is referred to \( \epsilon \). Need more information here.
can be covered as:

\[ q = \{ N(\mu', \sigma') : \mu' \in [-1, 1], \sigma' \in [\frac{1}{10}, 10], \text{ and } \mu' \text{ and } \sigma' \text{ are multiples of } \epsilon/100 \} \]

Obviously the covering has a polynomial dependence on \( \epsilon \).

In this case, cover size is \( O(1/\epsilon^2) \) and sample complexity is \( O\left(\frac{\log(1/\epsilon)}{\epsilon^2}\right) \). This example may not be useful in practice, but it is very useful from a theoretical standpoint to understand how tournament works with cover in learning.

4.2 Example 2

\[ \mathcal{C} = \{ \text{mixture of } \leq k \text{ Gaussian distributions of previous form} \} \]

And distribution belongs to class \( \mathcal{C} \) can be parameterized by \( \mu_1, \mu_2, ..., \mu_k, \sigma_1, \sigma_2, ..., \sigma_k, \) and \( w_1, w_2, ..., w_k \) where \( w_i \) represents the weight of the \( i^{th} \) Gaussian distribution.

By discretizing these parameters we can get a cover size of \( O((k/\epsilon)^2 k) \) with a sample complexity equals to \( O\left(\frac{k\log(k)/\epsilon^2}{\epsilon^2}\right) \). It is possible for us to apply the tournament algorithm after covering \( \mathcal{C} \) with these \( \epsilon \) balls.

This solution to the problem may not be efficient, but definitely is a way to solve the problem.

Notice: If we try to learn the parameters instead of the distribution itself, we may encounter some problem because \( p \) can have very similar parameters with a far distribution \( q_i \).

4.3 Miscellaneous

In practice, how to get the cover is usually non-trivial.

And the metric dimension has a relation relation with the covers:

\[
\text{metric dimension}(S) = \lim_{\epsilon \to 0} \left( \frac{\text{minimum number of balls in } \epsilon\text{-cover of } S}{\log(1/\epsilon)} \right)
\]

5 Another Explanation

Following is another way of talking about learning distribution on \([n]\).

5.1 Problem

Define \( \hat{p} \) as the empirical distribution of \( p \) we estimated from samples. We want to show that with high probability \( d_{tv}(p, \hat{p}) \) is small.
5.2 Analysis

To show that $d_{tv}(p, \hat{p}) < \epsilon$, we need to cover all $2^n$ $A$’s. By definition we have

$$d_{tv}(p, \hat{p}) = \sup_A |P(A) - \hat{P}(A)|$$

We want to take enough samples for each $A$ such that $|P(A) - \hat{P}(A)| < \epsilon$ with a probability $\geq 1 - (2^{-n}/4)$. Therefore we have a sample complexity of $O\left(\frac{\lg(2^n)}{\epsilon^2}\right)$, which is the same complexity we proved in previous sections.

6 Introduction to VC-dimension

With structured distributions, maybe we don’t need to consider all $A$’s. To apply the tournament algorithm, we only need $\hat{p}(A_{ij})$ to be accurate.

The following example is only covered until setup. Refer to lecture note on 01/30 for the full content. The following example also serves as an introduction to VC-dimension, which was not fully covered in this lecture because of limited time we had.

6.1 Simple Example

6.1.1 Problem

Consider the following class of distribution

$$\mathcal{C} = \{N(\mu, I), \text{Gaussian in } \mathbb{R}^d \text{ with identity covariance matrix}\}$$

In this case, $A_{ij}$ are half-spaces that separates all $q$ distributions. Half-spaces are infinite set with nice structures which we can take advantage of.

A natural question we may ask in this case is that: How many samples do we need until $|P(A) - \hat{P}(A)| < \epsilon$ for any half-space.

6.1.2 Setup

We are going to restrict the $L1$ distance by defining a new metric $\mathcal{A}$ such that

$$|p - q|_{\mathcal{A}} = \sup_{A \in \mathcal{A}} |p(A) - q(A)| < d_{tv}(p, q)$$

The intuition behind such setting is that we need to compare less sets if each set covers a larger part of $\mathcal{C}$. 
6.1.3 Double Sampling Argument

\(|p - \hat{p}|_A\) is small with high probability \(\iff\) \(|\hat{p} - \hat{p}'|_A\) is small with high probability.

**Proof.** We prove the above statement by proving the following two statements.

- Obviously, \(|p - \hat{p}|_A\) is small with high probability implies \(|\hat{p} - \hat{p}'|_A\) is small with high probability:
  
  By triangle inequality,
  
  \[|\hat{p} - \hat{p}'|_A \leq |p - \hat{p}|_A + |p - \hat{p}'|_A\]

- If \(|p - \hat{p}|_A\) is not small:
  
  - Assume \(|p - \hat{p}|_A > \epsilon\). Then for some \(A \in \mathcal{A}\) such that \(|p(A) - \hat{p}(A)| > \epsilon\).
  
  - Suppose we take enough samples with sample size \(>> 1/\epsilon^2\), which is always sufficient to learn a distribution. Then with high probability we have
    
    \[|p(A) - \hat{p}(A)| < \epsilon/2\]

    If so,
    
    \[|\hat{p}'(A) - \hat{p}(A)| > \epsilon/2\]
    
    \[\Rightarrow |\hat{p}' - \hat{p}| > \epsilon/2\]

If we take \(\Omega(1/\epsilon^2)\) samples, then

\[Pr(|\hat{p} - \hat{p}'|_A > \epsilon/2) \geq \frac{1}{2}Pr(|p - \hat{p}|_A > \epsilon)\]

In other words, \(\hat{p}\) take samples \(X_1, X_2, \ldots, X_n\) and \(\hat{p}'\) takes samples \(Y_1, Y_2, \ldots, Y_n\), and

\[
\frac{\# \{X_i \in A\} - \# \{Y_i \in A\}}{n}
\]

is small.

In order to eliminate the mild relation between \(X_i\)'s and \(Y_i\)'s, we pick \(n\) pairs of samples \(a_i, b_i\) and flip a fair coin to decide how to assign the values when sampling from the distribution. We assign \(a_i\) to \(X_i\) if the coin is head, and \(a_i\) to \(Y_i\) if the coin is tail (or the other way around).

Such procedure ensured that these samples are fixed and all independent from each other because they have a even chance to be assign \(a_i\) or \(b_i\).