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

Instructor: Jayadev Acharya — Lecture #5
Scribe: David Lee — 8th September, 2016

The topics we covered in class today are:

- Learning Tree Graphical Models
- Covering Numbers
- Mixtures of Gaussians

1 Learning Tree Graphical Models

1.1 Chow-Liu Algorithm

For n different people, and corresponding random variables $\{X_1, X_2,...,X_n\}$, we estimate the mutual information between random variables $\hat{I}(X_i; X_j)$ and look at the distributions over n samples. We know that the mutual information of two random variables is defined as

$$I(X_i; X_j) = -H(X_i, X_j) + H(X_i) + H(X_j) \tag{1}$$

$\hat{I}(X_i; X_j)$ gives the weight of a possible edge between X_i and X_j. From this information, we can produce a max-weight spanning tree that maximizes the data likelihood using greedy algorithms such as Kruskal's or Prim's algorithms.

2 Covering Numbers

We saw in Learning Distributions that for a collection of distributions $\mathcal{P}$ with radius ε, the smallest cover is N_ε for the total variation distance, and the minimum number of samples needed is $\frac{log(N_\varepsilon)}{\varepsilon^2}$.

For a collection of distributions $\{P_1, P_2,...,P_n\}$, with $P \in \Delta_k$, we want $d_{TV}(P, P_i) < \varepsilon$ for some i. Then for a $P = (p(1), p(2), ..., p(i))$, where $p(i) = j \cdot \frac{\varepsilon}{k}$, and p consists of $\frac{k}{\varepsilon}$ elements. Then, the covering number $N_c < (\frac{k}{\varepsilon})^k$. Then for any such P, the error distance for some $p(i)$ is $\pm \frac{\varepsilon}{2k}$ for a total error of $\pm \frac{\varepsilon}{2k} \cdot k = \varepsilon$. For such a distribution, $\frac{k \cdot log(\frac{k}{\varepsilon})}{\varepsilon^2}$ samples are sufficient.

Then suppose we know the collection of distributions $P_1, P_2,...,P_n$, and there is an unknown distribution P consisting of random variables $X_1, X_2,...,X_n$. Our goal is to then find a P_i in our known collection of distributions such that $d_{TV}(\mathcal{P}, P_i) < c \cdot min_j d_{TV}(\mathcal{P}, P_j) + O(\varepsilon)$. To solve this we will take a player vs. player type approach, where we compare the output the distribution it matches closest to. For example, consider two known distributions P_1 and P_2. Then we define A_{12} as

$$A_{12} = \{x \in X : P_1(x) - P_2(x)\} \tag{2}$$

Then for $X_1, ..., X_n$,

$$\mu(A) \triangleq \frac{|X_i \epsilon A|}{n} \tag{3}$$

Using μ, we compute $\mu(A_{12}$ which will give us an output of two different possibilities: $P_1(A_{12})$ or $P_2(A_{12}$. We compare the output to each distribution P_1 and P_2, and the distribution that matches closest to the output will be the distribution we will model as the unknown distribution. We can apply this to any number of known distributions, putting distributions against each other until the closest matching distribution is found.
To generalize, consider an underlying distribution Q^*. Then, for $\frac{log(N)}{\varepsilon^2}$ samples,

$$d_{TV}(P, Q^*) < c \cdot \Delta_{min} + O(\varepsilon) \tag{4}$$

where $\Delta_{min} = min_j d_{TV}(P, P_j)$.

3 Mixtures of Gaussians

Consider a collection of Gaussian distributions $\omega_1, \omega_2, ..., \omega_k$ where $\sum_{i=1}^{k} \omega_i = 1$. Each ω_i has a mean μ_i and a variance σ_i^2. Then a mixture of Gaussians is a combination of these normal Gaussians, which we can learn through 3 notions of learning:

- Proper Learning
- Improper Learning
- Parameter Learning

The main concept of learning mixtures of Gaussians is to find such a distribution where the total variation distance between the underlying distribution and some known or unknown distribution is minimal. We will only cover proper and improper learning in this section. For proper learning of mixtures of Gaussians, the underlying distribution is in a known mixture of Gaussians, with a max error of $\pm 2\varepsilon$. For improper learning, the underlying distribution is compared to an estimated distribution, not necessarily a mixture of Gaussians, with a max error of $\pm\varepsilon$.
The number of samples for both proper and improper learning are the same. The properties of proper and improper learning are shown in the table below:

	Samples	Time
improper	$\frac{k}{\varepsilon^2} log \frac{k}{\varepsilon}$	$\frac{k}{\varepsilon^2} log \frac{k}{\varepsilon}$
proper	$\frac{k}{\varepsilon^2} log \frac{k}{\varepsilon}$	$\frac{1}{\varepsilon}^{3k \pm 1}$

The time for improper learning can be expressed as a piecewise polynomial. For $X_1, ..., X_n$, we can find $\omega_1, ..., \omega_k$, $\mu_1, ..., \mu_k$ and $\sigma_1^2, ..., \sigma_k^2$ according to the table below:

$\omega_1, ..., \omega_k$	$\{\frac{\varepsilon}{k}, \frac{2\varepsilon}{k}, ..., 1\}$
$\mu_1, ..., \mu_k$	$\{X_1, ..., X_n\}$
$\sigma_1^2, ..., \sigma_k^2$	$\{X_i - X_j\}^2 \ \forall i, j$

The idea for setting these parameters according the table is that for $X_1, ..., X_n$, some will equate to the mean of a normal Gaussian distribution, and using this idea, the variance can be found by subtracting their values and squaring the result for all $X_1, ..., X_n$. Then for $(\frac{k}{\varepsilon} \cdot n)^k$ distributions, one of them will be close to $O(\varepsilon)$.