CS362 Algorithmic Frontiers April 21, 2014

Lecture 7

Lecturer: Gregory Valiant *Scribe: Katelyn Gao*

1 Topic

The next series of lectures will discuss finding structure in distributions. A classic problem in this area is finding clusters in data, which is NP-hard in the worst case. Therefore, we will assume that the data come from a nice "clustered" distribution, such as a mixture of Gaussians.

2 Gaussian Mixture Model

We define a Gaussian Mixture Model (GMM) with k components as follows:

Definition 1 *$G(\{N1, ..., N_k, w_1, ..., w_k\}) = \sum_i w_i N_i$, where N_i is normally distributed with mean μ_i and variance Σ_i. That is, for $i = 1, ..., k$, each sample is drawn from N_i with probability w_i.*

There are four possible learning problems:

- Parameter recovery of w_i, μ_i, Σ_i for $i = 1, ..., k$.
- "proper learning"/density estimation: Recover another GMM G' such that $G' \approx G$
- "improper learning"/density estimation: Recover any distribution D such that $D \approx G'$
  - Well studied in 1D, i.e. using histograms
- Clustering: label data such that samples from the same component have the same label

The first is "hardest" problem, in that if one accurately recovers approximations of the true parameters, then one can trivially perform the other tasks (although one can only hope to perform clustering up to the overlap in the densities of the components).

2.1 Dasgupta '99

The following theorem is motivated by high dimensional clustering. It basically says that in a GMM, if the distances between clusters are at least polynomial in the dimension of the data, then we can efficiently cluster samples and recover the cluster parameters.

Theorem 2 *If $G = (\{N_i, w_i\}), N_i \sim N(\mu_i, I_d)$, and $\forall i \neq j, \|\mu_i - \mu_j\|_2 > d^{1/4} \log^2 d$, then we can efficiently cluster samples.*

Proof For $X \sim N(0, I_d)$, as $d \to \infty$, wp $1, \|X\|_2 = \sqrt{d} \pm d^{1/4} \log d$. Therefore, we have the following.

Lemma 3 *For $X, Y \sim N_i$, with high probability $\|X - Y\|_2 = \sqrt{2d} \pm d^{1/4} \log d$. For $X \sim N_i, Y \sim N_j, i \neq j$, with high probability $\|X - Y\|_2 = \sqrt{2d + \sqrt{d} polylog(d)} \pm d^{1/4} \log d = \sqrt{2d} + d^{1/4} polylog(d) \pm d^{1/4} \log d$.*

Hence, with high probability the following algorithm will work: $\forall X_i$, label all X_j such that $\|X_i - X_j\|_2 \leq \sqrt{2d} + d^{1/4} \log d$ with the same cluster. ■

3 Method of Moments

The Method of Moments dates back to Pearson in 1890. For illustration, consider a random variable X, which is a GMM with 2 components in 1D: $G = \{w_1, w_2, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2\}$. We have

$$\mathrm{E}_{x \sim G} X^r = w_1 \mathrm{E}_{x \sim N(\mu_1, \sigma_1^2)} X^r + w_2 \mathrm{E}_{x \sim N(\mu_2, \sigma_2^2)} X^r$$

Therefore, the rth moment of X is a rth degree polynomial in the parameters.

The idea of Method of Moments is to find parameters whose corresponding moments match the empirical moments. For a GMM, this reduces to solving a system of polynomial equations.

Fact 4 *To estimate the rth moment of a Gaussian to a factor of $\pm\epsilon$, we need at least $1/\epsilon^{2r}$ samples.*

The problem is further exacerbated in high dimensions, since we must estimate tensors instead of scalars. The moral of the story is to avoid high moments if possible.

4 Fitting GMMs using Method of Moments

Fact 5 *Projections of high dimensional Gaussians are Gaussian, so projections of high dimensional GMMs are GMMs. Moreover, given a fixed projection subspace, the parameters of a projected GMM are known functions of the parameters of the original GMM.*

From the above fact, we see that to avoid high moments in large dimensions, we can reduce the problem of learning GMMs to a sequence of $O(d^2)$ 1D problems via random projections. So it suffices to consider the problem of learning a GMM in 1D, with k components. The following lemma shows that it is possible.

Lemma 6 *Any two GMMs G, G', each with k components, differ in at least one of the first $2(2k-1)$ moments.*

This lemma relies on the following basic claim about convolution with a Gaussian (which is typically interpreted via the heat equation for the diffusion of heat along a 1-dimensional rod):

Claim 7 *For any analytic function $g : \mathbb{R} \to \mathbb{R}$, $\sigma^2 > 0$, $g * N(0, \sigma^2)$ has at most as many zero crossings as g.*

We now prove Lemma **??**, modulo a little hand-waving about delta functions—to make it rigorous, one should imagine the delta functions being Gaussians with vanishingly small variances. Define a function $f = \sum_{i=1}^{k'} \alpha_i N(\mu_i, \sigma_i^2)$. We claim that f has at most $2(k'-1)$ zero crossings. This can be proved inductively, beginning with the case $k' = 1$. For general k', consider the function f but with a component of minimal variance removed, and all other components with their variance reduced by the variance of the minimal component. By our inductive hypothesis, this function has at most $2(k-2)$ zeros, and the addition of the final component (as a delta function) can increase the number of zero-crossings by at most 2, and the convolution by the intended variance of this component can only reduce this number, by the above claim.

Given the above, for f that is the difference in densities of two GMMs with each having at most k components, f must have at most $2(2k-1)$ zeros, hence there exists a polynomial $P(x)$ with degree at most $2(2k-1)$ that agrees with f in sign everywhere, so we have $\int P(x)f(x) > 0$. Then, at least one of the first $2(2k-1)$ moments is different for the two GMMs. With a bit of work, one can show that this difference must be significant if the means, or variances of the components of either GMM differ significantly.

Note that the above approach seems to require the number of samples to be exponential in k, which is unfortunate, though provably necessary, unless we make additional assumptions on the GMMs.