EE 381V: Large Scale Optimization **Spring 2013**

Lecture 9 — February 12

Lecturer: Caramanis & Sanghavi *Scribe: Ethan R. Elenberg*

9.1 Review

A Gaussian Mixture Model (GMM) is a probability distribution with the following probability density function:

$$f(\boldsymbol{x}) = \sum_{i=1}^{k} w_i \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i). \tag{9.1}$$

Here, the mixing weights $\{w_i\}$ are selected such that their sum $\Sigma_i w_i = 1$. One of k multivariate Gaussian distributions is chosen based on these weights, and then the point is sampled from that distribution. We will consider clustering GMM's for the isotropic case in which $\Sigma_i = \sigma_i^2 I$. Last time we saw the following:

- High dimensionality is a problem.
- Projecting onto a random, low-dimensional subspace does not solve the problem.
- Projecting onto a subspace passing through $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ does solve the problem.

We also introduced the following algorithm for clustering n-dimensional data from a mixture model:

Algorithm 1 Spectral Clustering for GMM's [1]

1: Form a sample matrix $A \in \mathbb{R}^{m \times n}$, where m is the number of points.
2: Find $\hat{V} = \arg\max_{V:\, \dim(V)=r} \|\text{Proj}_V(A)\|^2$ via truncated SVD:

$$A \approx \hat{U}_r \hat{\Sigma}_r \hat{V}_r^T$$

3: Project every point onto $\hat{V}$.
4: Do something simple in reduced dimensions, i.e. distance based clustering.

9.2 Expected Optimal Subspace for Clustering

Recall that in the 1-dimensional case, if $x \sim \mathcal{N}(\mu, \sigma^2 I)$, then

$$\hat{v} = \underset{v:\,\dim(v)=1}{\arg\max}\; E[\|\text{Proj}_v(x)\|^2] = \frac{\mu}{\|\mu\|}. \tag{9.2}$$

Now we will prove the extension to r dimensions that was argued last lecture. Intuitively, there are many r-dimensional subspaces which pass through each $\boldsymbol{\mu}_i$ individually, but a subspace which passes through all of the means will be jointly optimal overall.

Formally, consider the "expected matrix" $E[A]$:

$$E[A] = \begin{bmatrix} - & \boldsymbol{\mu}_1 & - \\ - & \boldsymbol{\mu}_1 & - \\ - & \boldsymbol{\mu}_1 & - \\ & \vdots & \\ - & \boldsymbol{\mu}_k & - \\ - & \boldsymbol{\mu}_k & - \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ w_1 m \\ \\ w_k m \\ \\ \end{matrix} \tag{9.3}$$

Each mean is repeated a number of times proportional to its mixing weight. This may be interpreted as the data matrix A in the case of no randomness (each point is one of the deterministic means). The following theorem rewrites the expected projection in terms of $E[A]$:

Theorem 9.1. *Let $A \in \mathbb{R}^{m\times n}$ be a sample matrix corresponding to a GMM with k distributions $\mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I)$, $1 \le i \le k$. For any V of dimension r,*

$$E[\|\text{Proj}_V A\|^2] = \|\text{Proj}_V E[A]\|^2 + m\sum_{i=1}^{k} w_i \sigma_i^2 r. \tag{9.4}$$

Interpretation: This expectation may be split into a deterministic function of V and a random component which only depends on r (assumed to be fixed). Thus, the V which maximizes the left hand side is the one which maximizes the first term of the right hand side.

Comment: The "best" V (i.e. with largest projection) passes through $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$.

Proof:

$$\begin{aligned} E[\|\text{Proj}_V A\|^2] &= \sum_{i=1}^{m} E[\|\text{Proj}_V A_i\|^2] = \sum_{i=1}^{m}\sum_{l=1}^{k} E[\|\text{Proj}_V A_i\|^2 \,|i \in F_l]\,\mathbf{Pr}\,(i \in F_l) \\ &= \sum_{i=1}^{m}\sum_{l=1}^{k} w_l\, E[\|\text{Proj}_V A_i\|^2 \,|i \in F_l] \end{aligned}$$

For a single point, project using an orthonormal basis $\{v_j\}$ of V,

$$\text{Proj}_V A_i = \sum_{j=1}^{r} \langle A_i, v_j \rangle v_j$$

$$\Rightarrow E[\|\text{Proj}_V A_i\|^2] = \sum_{j=1}^{r} E[\langle A_i, v_j \rangle^2]$$

For every j,

$$\begin{aligned} E[\langle A_i, v_j \rangle^2] &= E[\langle \boldsymbol{\mu}_i + \boldsymbol{x}, v_j \rangle^2] = E[\langle \boldsymbol{\mu}_i, v_j \rangle^2 + \langle \boldsymbol{x}, v_j \rangle^2 + 2\langle \boldsymbol{\mu}_i, v_j \rangle^2 \langle \boldsymbol{x}, v_j \rangle^2] \\ &= \langle \boldsymbol{\mu}_i, v_j \rangle^2 + \sigma_i^2, \end{aligned}$$

where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \sigma_i^2 I)$. The cross terms are zero because $\boldsymbol{x}$ is zero mean. Now summing this deterministic result over j,

$$E[\|\text{Proj}_V A_i\|^2] = \|\text{Proj}_V \boldsymbol{\mu}_i\|^2 + r\sigma_i^2.$$

Summing over all points i and noting that $i \in F_l \Rightarrow \boldsymbol{\mu}_i = \boldsymbol{\mu}_l$, $\sigma_i^2 = \sigma_l^2$,

$$\begin{aligned} \Rightarrow E[\|\text{Proj}_V A\|^2] &= \sum_{i=1}^{m} \sum_{l=1}^{k} w_l \left(\|\text{Proj}_V \boldsymbol{\mu}_l\|^2 + r\sigma_l^2 \right) = \sum_{l=1}^{k} m w_l \|\text{Proj}_V \boldsymbol{\mu}_l\|^2 + m \sum_{l=1}^{k} w_l r \sigma_l^2 \\ &= \|\text{Proj}_V E[A]\|^2 + m \sum_{i=1}^{k} w_i \sigma_i^2 r. \end{aligned}$$

□

9.3 Concentration

In Algorithm 1, we cluster by projecting onto the subspace $\hat{V} = \arg\max_V \|\text{Proj}_V A\|^2$. However, Theorem 9.1 only shows a result for $V^* = \arg\max_V E[\|\text{Proj}_V A\|^2]$. Now we need to show that $\hat{V} \approx V^*$ with high probability for m large enough. We prove this concentration in 2 steps:

(a) For a given, fixed V, $\|\text{Proj}_V A\|^2 \approx E[\|\text{Proj}_V A\|^2]$.

(b) Step (a) is true for every V simultaneously with high probability.

Lemma 9.2 (Step (a)). *For a given, fixed V,*

$$\textbf{Pr}\left(\|\text{Proj}_V A\|^2 \geq (1+\epsilon) E[\|\text{Proj}_V A\|^2] \right) < k e^{-\epsilon^2 m r / 8}. \tag{9.5}$$

Similarly, $$\textbf{Pr}\left(\|\text{Proj}_V A\|^2 \leq (1-\epsilon) E[\|\text{Proj}_V A\|^2] \right) < k e^{-\epsilon^2 m r / 8}. \tag{9.6}$$

Proof: For V fixed, simply find an ensemble basis and use concentration results on each Gaussian distribution separately. The process is simple but lengthy; refer to [1] for more details. □

9.3.1 Covering Argument

Proving Step (b) is somewhat more involved. We cannot use a union bound over all V's because such a bound loses all meaning when taken over uncountably many subspaces. Instead of directly proving that the Step (a) holds with high probability for uncountably many V's, we use a covering argument: a proof technique revisited in the sequel. We modify Equations 9.5 and 9.6 by adding another term to the inequality and showing that this new result holds for a finite collection of subspaces as well as for all subspaces contained in a neighborhood around each subspace:

Lemma 9.3 (Step (b)). *For any* $1 > \epsilon > 0$ *and* $0 < \alpha < \frac{1}{\sqrt{n}}$,

$$\mathbf{Pr}\left(\exists\, V \text{ s.t. } \|\text{Proj}_V A\|^2 > (1+\epsilon)E[\|\text{Proj}_V A\|^2] + 6r\sqrt{n}\alpha E[\|A\|^2]\right) < \left(\frac{2}{\alpha}\right)^{rn} k e^{-\epsilon^2 mr/8}. \tag{9.7}$$

Proof: The sketch in 2 dimensions is as follows: consider a finite set of vectors, whose elements have minimum separation distance no more than α, with cardinality inversely proportional to α. For these vectors, the bounds in Step (a) hold without the "fudge factor" $6r\sqrt{n}\alpha E[\|A\|^2]$, and a union bound results in the term $\left(\frac{2}{\alpha}\right)^{rn}$. Vectors outside the set are sufficiently close to an element of the set to bound them with the fudge factor. □

9.3.2 Number of Points

Now we find an m such that the bound in Lemma 9.3 is small and $\hat{V} \approx V^*$. Taking a natural log results in the inequality $rn \log \frac{2}{\alpha} < \frac{\epsilon^2 mr}{8}$. Recalling that α is a negative power of n, we conclude that m should scale $O(n \log n)$.

9.4 Nonspherical GMM

This spectral clustering algorithm works well for the spherical GMM case, but it fails miserably for the "2 pancakes" problem. Next class, we will present a modified algorithm that obtains an appropriate subspace for clustering in this setting.

Bibliography

[1] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *J. Comp. Sys. Sci.*, 68(2):841-860, 2004.