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

Lecture 8 — Feb 7

Lecturer: Caramanis & Sanghavi *Scribe: Hsiang-Fu Yu*

8.1 Introduction

We have learned some approaches of Dimension Reduction (DR), such as LSH for nearest neighbor search and spectral clustering for a given similarity graph (or matrix). In the lecture today, we are going to introduce spectral clustering for Gaussian Mixture Models (GMM).

8.2 Gaussian Mixture Models

A Gaussian mixture model is a distribution with the probability density function defined as follows.

$$P(\boldsymbol{x}) = \sum_{i=1}^{k} w_i \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i), \tag{8.1}$$

where $\boldsymbol{x} \in \mathcal{R}^n$ is a sampled point, $w_i \geq 0 \ \forall i$, $\sum_{i=1}^{k} w_i = 1$, and $\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$ is a multivariate Gaussian distribution characterized by the mean vector $\boldsymbol{\mu}_i$ and the covariance matrix Σ_i.

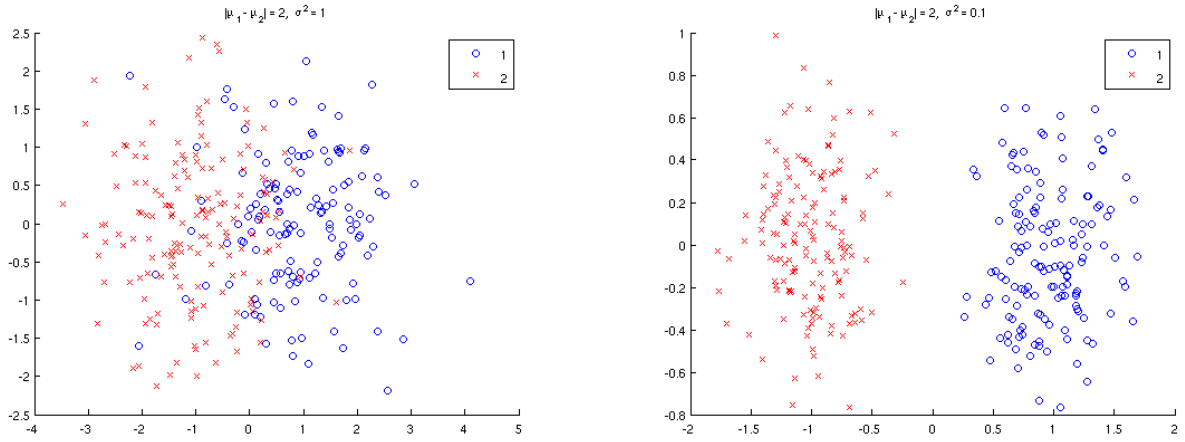
Sampling Process. We first select a index $i \in \{1, \ldots, k\}$, where each i is selected with the probability w_i, then sample a point $\boldsymbol{x}$ from $\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$.

8.2.1 Clustering Problem

Given m points (without index label) sampled from a GMM, where only the parameter k is known, we want to find the correct index label for each point.

The difficulty of this problem depends on the parameters of the underlying GMM (i.e., $\{\boldsymbol{\mu}_i\}$ and $\{\Sigma_i\}$). Let's look at a simple example in Figure 8.1, where we consider a simplified GMM with $k = n = 2$, $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| = 2$, and $\Sigma_1 = \Sigma_2 = \sigma^2 I_2$. Figure 8.1a shows the result for $\sigma^2 = 1$, while Figure 8.1b shows the result for $\sigma^2 = 0.1$. Obviously, the clustering problem for $\sigma^2 = 0.1$, where points generated from different Gaussian distributions do not overlap, is easier than clustering for $\sigma^2 = 1$, where points are highly mixed. In general, for any two Gaussian $\mathcal{N}_i$ and $\mathcal{N}_j$, if $E[\|X^{(i)} - \boldsymbol{\mu}_i\|]$ and $E[\|X^{(j)} - \boldsymbol{\mu}_j\|]$ are much smaller than $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$, where X^i denotes a random data point sampled from $\mathcal{N}_i$, then clustering problem becomes easier. Therefore, the distance between $\{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| : i \neq j\}$ and $\{E\|X^i - \boldsymbol{\mu}_i\|\}$ is the key to determine the difficulty of the clustering problem.

We state a useful lemma for estimate $E[\|X - \boldsymbol{\mu}\|]$ for a multivariate Gaussian:



(a) $\sigma^2 = 1$ (b) $\sigma^2 = 0.1$

Figure 8.1: Clustering Difficulty for different σ^2

Lemma 8.1. *For an n-dimensional Gaussian distribution $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$,*

$$E[\|X - \boldsymbol{\mu}\|^2] = \sum_{i=1}^{n} \sigma_{ii}^2,$$

where σ_{ii} is the i-th entry in the diagonal of Σ.

Proof: By definition, $X = \boldsymbol{\mu} + Y$, where $Y \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Thus,

$$E[\|X - \boldsymbol{\mu}\|^2] = E[\|Y\|^2] = E[\sum_{i=1}^{n} Y_i^2] = \sum_{i=1}^{n} E[Y_i^2] = \sum_{i=1}^{n} \sigma_{ii}^2.$$

□

8.3 Settings

In this lecture, we consider the clustering problem for a simplified version of GMM:

- the dimension n is large,
- the covariance matrix for each Gaussian distributions is just a diagonal matrix $\Sigma_i = \sigma_i^2 I_n$,.
- $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ is considered a constant which is independent of the dimension n.

Based on Lemma 8.1, we have the following Corollary to measure the difficulty of clustering problems under our setting.

Corollary 8.2. *For an n-dimensional Gaussian distribution $X \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$,*

$$E[\|X - \boldsymbol{\mu}\|] = \sigma\sqrt{n}.$$

As a result, to obtain a good clustering result, distance-based methods such as K-means require that

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| > C \max\{\sigma_i, \sigma_j\}\sqrt{n} \tag{8.2}$$

holds for $i \neq j$, where C is a constant. As the RHS of (8.2) is linear to $\sqrt{n}$, distance-based methods will fail when n is large.

Does this high-dimensional clustering problem become easier when $\{\boldsymbol{\mu}_i\}$ is also given? Consider a simple case where $k = 2$ and $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are known. We can simply project all data points on the line connecting $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$,[1] then

$$E[\|\operatorname{Proj}(X^{(i)}) - \operatorname{Proj}(\boldsymbol{\mu}_i)\|] = \sigma_i, \ i = 1, 2.$$

As a result, as long as $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\| > C \max\{\sigma_1, \sigma_2\}$, distance-based methods can work well on the projected data because the RHS is independent of n. In general, if $\{\boldsymbol{\mu}_i\}$ are told, and $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| > C \max\{\sigma_i, \sigma_j\}$ for each pair of (i, j), the clustering problem becomes easier when we apply appropriate projection.

8.4 Two Ideas for Projection

However, in real-world application, $\{\boldsymbol{\mu}_i\}$ are usually unknown. Here we try two ideas to find an appropriate projection.

8.4.1 Idea I - Random Projection

The first idea is projecting the data onto a random r-dimensional subspace V, where $n \gg r > k$. As the dimension becomes r, $\sigma\sqrt{n}$ becomes $\sigma\sqrt{r}$, which is a good thing as the RHS in (8.2) is reduced. However, by the Johnson-Lindendstrauss lemma [1, 2],

$$E[\|\operatorname{Proj}_V(\boldsymbol{\mu}_i) - \operatorname{Proj}_V(\boldsymbol{\mu}_j)\|^2] = \frac{r}{n}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2,$$

which means that the LHS in (8.2) is also reduced by random projection. This means that if distance-based methods are not able to cluster the original data well, the projected data cannot be well-clustered as the difficulty remains after the random projection.

[1] We can assume $\boldsymbol{\mu}_2 = c\boldsymbol{\mu}_1$ for some constant c to make the line an one-dimensional subspace.

Algorithm 1 Spectral Clustering for Gaussian Mixture Models

1: Form the $m \times n$ sample matrix A, where each row is data point.
2: Calculate the truncated SVD for A with rank r:

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

3: Form the projected $m \times r$ sample matrix $A' = \hat{U}_r \hat{\Sigma}$.
4: Run an elementary clustering algorithm on A'.

8.4.2 Idea II - Projection covering $\text{Span}\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$

The reason why random projection does not work is that it also reduces the distance between $\{\boldsymbol{\mu}_i\}$. If we knew a r-dimensional subspace U that contains $\text{Span}\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$, then after a projection onto U, the RHS in (8.2) is reduced, while the distance between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ (i.e., the LHS in (8.2)) remains unchanged:

$$\begin{aligned} &\|\operatorname{Proj}_U(\boldsymbol{\mu}_i) - \operatorname{Proj}_U(\boldsymbol{\mu}_j)\| = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|, \ \forall i, j \\ &E[\|\operatorname{Proj}_U(X^i) - \operatorname{Proj}_U(\boldsymbol{\mu}_i)\|] = \sigma_i \sqrt{r}, \ \forall i. \end{aligned}$$

As a result, clustering projected data becomes easier for distance-based methods. Next we discuss how to find the desired subspace U.

8.5 Spectral Clustering for GMMs

Spectral clustering for GMMs is an approach to find/approximate the projection "U" described in Section 8.4.2.

Intuition. Assume that X is generated from a single Gaussian $\sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$ and consider the optimization problem:

$$\arg\max_{\|\boldsymbol{v}\|=1} E[< X, \boldsymbol{v} >^2]. \tag{8.3}$$

Recall that $X = \boldsymbol{\mu} + Y$, $Y \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, thus $< X, \boldsymbol{v} >=< \boldsymbol{\mu}, \boldsymbol{v} > + < Y, \boldsymbol{v} >$. As $E[< Y, \boldsymbol{v} >]$ is a constant for all $\boldsymbol{v}$,

$$\begin{aligned} &\arg\max_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} E[< X, \boldsymbol{v} >^2] \\ =&\arg\max_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} E[< X, \boldsymbol{v} >] \\ =&\arg\max_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} E[< \boldsymbol{\mu}, \boldsymbol{v} >] + E[< Y, \boldsymbol{v} >] \\ =&\arg\max_{\boldsymbol{v}:\|\boldsymbol{v}\|=1} E[< \boldsymbol{\mu}, \boldsymbol{v} >] + \text{constant} \\ =&\boldsymbol{\mu}. \end{aligned}$$

The optimal solution for (8.3) is just $\boldsymbol{\mu}$. Similarly, for a GMM with $k > 1$, we have

$$\arg \max_{V:dim(V)=k} E[\| \operatorname{Proj}_V X\|^2] = \operatorname{Span}\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}.$$

Thus, ideally, we can find a desired projection U through finding a projection which maximizing the expected length of projected data.

Given m data points sampled from a GMM following the setting in Section 8.3, the goal is to find a projection which maximizing the empirical expectation of length of the projected data.

$$\arg \max_{V:dim(V)=r} \frac{1}{m} \sum_{i=1}^{m} \| \operatorname{Proj}_V \boldsymbol{x}_i\|_2^2, \tag{8.4}$$

where $\boldsymbol{x}_i \in \mathcal{R}^n$ is the i-th data point. It can be analytically shown that the optimal projection is $\hat{V}_r^T$, the transpose of the matrix corresponding to the top-r right singular vectors of the $m \times n$ sample matrix A, where i-th row, $A_i = \boldsymbol{x}_i^T$, is the i-th data point. In particular, if we form the rank-r truncated SVD for A:

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

we have

$$\hat{V}_r^T = \arg \max_{V:dim(V)=r} \frac{1}{m} \sum_{i}^{m} \| \operatorname{Proj}_V A_i\|^2,$$

$$\hat{U}_r \hat{\Sigma}_r = \operatorname{Proj} A, \text{ the projected } m \times r \text{ sample matrix }.$$

As a result, we can conduct an elementary clustering algorithm on the projected sample matrix $\operatorname{Proj} A$. We describe the spectral algorithm for GMM in Algorithm 1.

Bibliography

[1] W. Johnson and J. Lindenstrauss, "Extensions of lipshitz mapping into hilbert space," in *Modern analysis and probability*, vol. 26, pp. 189–206, 1984.

[2] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, (New York, NY, USA), pp. 245–250, ACM, 2001.