University of Illinois Fall 2018

ECE 544NA: Pattern Recognition
Lecture 17: October 23

Lecturer: Alexander Schwing Scribe: Ankit Raj

1 Introduction

This lecture introduces another paradigm of statistical modeling, known as generative modeling. Particularly, the lecture talks about modeling the data using the Gaussian Mixture Models (GMMs) [1] and how they are related to the K-Means algorithm.

The following sections are organized as follows: Section 2 discusses about the standard generative modeling framework and how it is different from discriminative approach; Section 3 describes one of the simplest generative modeling technique using single Gaussian; Section 4 extends single Gaussian to multiple Gaussians case and provides details how they are similar to K-means clustering algorithm.

2 Generative Modeling

Generative modeling is a technique to learn the statistics of data. This can be used to generate new samples from the distribution of data used for training the model. Formally speaking, using data samples $x^{(i)}$ for , the model learns $p(x^{(i)}|\theta)$ where θ are the model parameters.

On the other hand, discriminative approach learns the statistics of labels or target variables, $y^{(i)}$ given data $x^{(i)}$ i.e. it learns $p(y^{(i)}|x^{(i)})$ for (data, target) pairs as $(x^{(i)}, y^{(i)})$.

An example below illustrates mathematical formulation of two algorithms, linear regression which follows a discriminative modeling approach and Gaussian model which is generative:

$$p(y^{(i)}|x^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T\phi(x^{(i)})^2\Big) \tag{1}$$

$$p(x^{(i)}|\mu,\sigma) = \mathcal{N}(x^{(i)}|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{1}{2\sigma^2}(x^{(i)} - \mu)^2\Big). \tag{2}$$

Important difference which can be noted from the above equations is, linear regression (Equation 1) is modeling the class labels $y^{(i)}$ given the data $x^{(i)}$ using the model parameters $\mathbf{w}$, whereas the Gaussian model (Equation 2) models the distribution of the data $x^{(i)}$ using the parameters μ and σ, also referred as θ or $\mathbf{w}$. Though it is sometimes ambiguous to what to call data or labels amongst $x^{(i)}$ and $y^{(i)}$ but generally, $x^{(i)}$ is referred as data and $y^{(i)}$ is referred as labels.

3 Single Gaussian Model

Equation 2 refers to the density function of Gaussian distributed data (with single Gaussian component).

Given a dataset $D = \{(x^{(i)})\}$, with $x^{(i)}$ being i.i.d i.e., independent and identically distributed, the likelihood is given by:

$$L(\mu, \sigma) = \prod_{i \in D} \mathcal{N}(x^{(i)}|\mu, \sigma) = \prod_{i \in D} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{1}{2\sigma^2}(x^{(i)} - \mu)^2 \Big). \tag{3}$$

3.1 Learning parameters

Parameters of the distribution, $(\theta = \mu, \sigma)$ are obtained by maximizing the above likelihood. Using the monotonic property of $log(.)$, maximizing $L(\mu, \sigma)$ is equivalent to minimizing $-log(L(\mu, \sigma))$. Hence, the program becomes:

$$\begin{aligned}
\min_{\mu,\sigma} -log(L(\mu, \sigma)) &:= -log \prod_{i \in D} p(x^{(i)}|\mu, \sigma) \\
&:= -log \prod_{i \in D} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{1}{2\sigma^2}(x^{(i)} - \mu)^2 \Big) \\
&:= \frac{N}{2} log(2\pi\sigma^2) + \sum_{i \in D} \frac{1}{2\sigma^2} \big(x^{(i)} - \mu\big)^2
\end{aligned} \tag{4}$$

where N is total number of samples in the dataset. To obtain the optimality condition, take derivatives of Equation 4 with respect to (w.r.t) the parameters, μ and σ and set them to zero.

Derivative w.r.t μ and σ is given by:

$$\begin{aligned}
\frac{\partial}{\partial \mu} : \frac{1}{\sigma^2} \sum_{i \in D} \big(x^{(i)} - \mu\big) &= 0 \\
\Rightarrow \quad \hat{\mu} &= \frac{1}{N} \sum_{i \in D} x^{(i)}
\end{aligned} \tag{5}$$

$$\begin{aligned}
\frac{\partial}{\partial \sigma} : \frac{-1}{\sigma^3} \sum_{i \in D} \big(x^{(i)} - \mu\big)^2 + \frac{N}{\sigma} &= 0 \\
\Rightarrow \quad \hat{\sigma}^2 &= \frac{1}{N} \sum_{i \in D} (x^{(i)} - \mu)^2
\end{aligned} \tag{6}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood estimates of μ and σ respectively. In Equation 6, μ is replaced by its estimate i.e. $\hat{\mu}$. $\hat{\mu}$ and $\hat{\sigma}$ are sample mean and sample variance respectively.

3.2 Issue with single Gaussian

Figure 1 clearly illustrates the issue with modeling the data using a single Gaussian. Left subplot has been modeled using only one Gaussian. It assigns the highest probability in the central region where there are very less data-points (not desirable), hence using single Gaussian doesn't provide flexibility to modeling even slightly complex data i.e. data with more than one modes.

Modeling using multiple Gaussians provide more flexibility as the right subplot shows modeling the data with 2 Gaussians. It can be clearly seen that it assigns higher probability to regions containing more data-points (desirable).

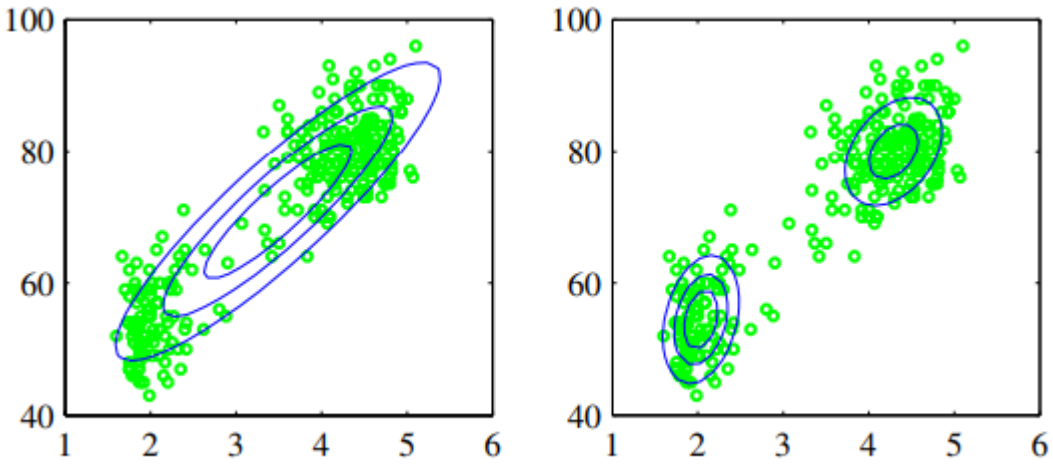

Figure 1: Modeling data using {Left} : single Gaussian, {Right} : double Gaussian

4 Mixture of Gaussians

Motivation for moving to mixture of Gaussians is very simple i.e. to be able to learn statistics of complex data and add flexibility in modeling. Data distribution for this case which is linear superposition of Gaussians is given by:

$$p(x^{(i)}|\pi,\mu,\sigma) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k) \tag{7}$$

with the following constraint:

$$\sum_{k=1}^{K} \pi_k = 1 \quad s.t. \quad \pi_k \geq 0$$

where $\theta = (\pi,\mu,\sigma)$ are the model parameters. K is the total number of Gaussian components used for modeling. π_k are also called as mixing coefficients as it decides the proportion from each Gaussian for data points. To learn the model parameters, the obvious approach is maximizing the log-likelihood or equivalently, minimizing negative log-likelihood (as done for single Gaussian case) with additional constraint of $\sum_{k=1}^{K} \pi_k = 1 \quad s.t. \quad \pi_k \geq 0$.

$$\begin{aligned} \min_{\pi,\mu,\sigma} -log(L(\pi,\mu,\sigma)) &:= -log \prod_{i \in D} p(x^{(i)}|\pi,\mu,\sigma) \\ &:= -\sum_{i \in D} log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k) \end{aligned} \tag{8}$$

Unlike the single Gaussian case, closed form solution of Equation 8 doesn't exist. However, it can still be optimized using gradient descent algorithm.

4.1 Introducing latent variables

4.1.1 Motivation

If we define a joint distribution over observed and latent variables, the corresponding distribution of the observed variables alone is obtained by marginalization. This allows relatively complex marginal distributions over observed variables to be expressed in terms of more tractable joint distributions over the expanded space of observed and latent variables. The introduction of latent variables thereby allows complicated distributions to be formed from simpler components. We will see this in the Gaussians mixture set-up.

4.1.2 Latent variables for GMM

Let z_{ik} be a binary auxiliary/latent variable and $z_{ik} \in \{0, 1\}$ with $\sum_{k=1}^{K} z_{ik} = 1 \ \forall$ i. With these latent variables, Expectation Maximization (EM) algorithm will be introduced to find the maximum likelihood of the parameters. By definition, latent variable is the identity of the Gaussian component that generated a given data point. Marginal for z_{ik} is given by:

$$\begin{aligned} p(z_{ik} = 1) &= \pi_k \\ p(\mathbf{z}_i) &= \prod_{k=1}^{K} \pi_k^{z_{ik}} \end{aligned} \tag{9}$$

where $\mathbf{z}_i = [z_{i1}, z_{i2}, ..., z_{iK}]^T$. Conditional for $x^{(i)}$ given $z_{ik} = 1$ is:

$$p(x^{(i)}|z_{ik} = 1) = \mathcal{N}(x^{(i)}|\mu_k, \sigma_k) \tag{10}$$

Since, $z_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} z_{ik} = 1$, conditional for $x^{(i)}$ given $\mathbf{z}_i$:

$$\begin{aligned} p(x^{(i)}|\mathbf{z}_i) &= \prod_{k=1}^{K} p(x^{(i)}|z_{ik} = 1)^{z_{ik}} \\ &= \prod_{k=1}^{K} \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)^{z_{ik}} \end{aligned} \tag{11}$$

The marginal for $x^{(i)}$using the latent variables can be computed as:

$$\begin{aligned} p(x^{(i)}|\pi, \mu, \sigma) &= \sum_{\mathbf{z}_i} p(x^{(i)}|\mathbf{z}_i) p(\mathbf{z}_i) \\ &= \sum_{\mathbf{z}_i} \prod_{k=1}^{K} \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)^{z_{ik}} \pi_k^{z_{ik}} \qquad (12) \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k) \qquad (13) \end{aligned}$$

Equation 13 has been obtained from the Equation 12 using the definition of latent variables i.e. $z_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} z_{ik} = 1$. As expected, marginal obtained using latent variable (Equation 13) is same as that of distribution defined in Equation 7.

Conditional probability of z_{ik} given $x^{(i)}$ (intuitively, probability that given $x^{(i)}$ belongs to the kth Gaussian component), also called as posterior is derived using Bayes rule.

$$\begin{aligned} r_{ik} &= p(z_{ik}=1|x^{(i)}) \\ &= \frac{p(z_{ik}=1)p(x^{(i)}|z_{ik}=1)}{p(x^{(i)})} \\ &= \frac{p(z_{ik}=1)p(x^{(i)}|z_{ik}=1)}{\sum_{\hat{k}=1}^{K} p(z_{i\hat{k}}=1)p(x^{(i)}|z_{i\hat{k}}=1)} \\ &= \frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^{(i)}|\mu_{\hat{k}},\sigma_{\hat{k}})} \end{aligned} \tag{14}$$

4.2 Learning Parameters

Using the equations obtained by introducing the latent variables, negative log-likelihood can be minimized. Recall that negative likelihood is given by (same as Equation 8):

$$\min_{\pi,\mu,\sigma} -log \prod_{i\in D} p(x^{(i)}|\pi,\mu,\sigma) = -\sum_{i\in D} log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k) \tag{15}$$
$$s.t. \quad \sum_{k=1}^{K} \pi_k = 1$$

To obtain the stationary points, take derivatives of Equation 15 with respect to (w.r.t) the parameters, μ_k and σ_k and set them to zero. Per-cluster weight for kth cluster across data samples, $N_k = \sum_{i\in D} r_{ik}$.

Derivative w.r.t μ_k is given by:

$$\begin{aligned} &\frac{\partial}{\partial \mu_k} : -\sum_{i\in D} \frac{\frac{\partial}{\partial \mu_k}\pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'},\sigma_{k'})} = 0 \\ \Rightarrow \quad &-\sum_{i\in D} \frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'},\sigma_{k'})} \Big(\frac{-1}{2\sigma_k^2}\big(x^{(i)}-\mu_k\big)\Big) = 0 \\ \Rightarrow \quad &-\sum_{i\in D} r_{ik} \Big(\frac{-1}{2\sigma_k^2}\big(x^{(i)}-\mu_k\big)\Big) = 0 \qquad (\textit{Using Equation 14}) \\ \Rightarrow \quad &\mu_k = \frac{1}{N_k}\sum_{i\in D} r_{ik} x^{(i)} \end{aligned} \tag{16}$$

Derivative w.r.t σ_k is given by:

$$\frac{\partial}{\partial \sigma_k} : -\sum_{i \in D} \frac{\frac{\partial}{\partial \sigma_k} \pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'}, \sigma_{k'})} = 0$$

$$\Rightarrow \quad -\sum_{i \in D} \pi_k \frac{\frac{\partial}{\partial \sigma_k} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\Big(-\frac{1}{2\sigma_k^2}(x^{(i)} - \mu_k)^2 \Big)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'}, \sigma_{k'})} = 0$$

$$\Rightarrow \quad \sum_{i \in D} \frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'}, \sigma_{k'})} \Big(\frac{1}{\sigma_k} - \frac{1}{\sigma_k^3} \big(x^{(i)} - \mu_k\big)^2 \Big) = 0$$

$$\Rightarrow \quad \sum_{i \in D} r_{ik} \Big(\frac{1}{\sigma_k} - \frac{1}{\sigma_k^3} \big(x^{(i)} - \mu_k\big)^2 \Big) = 0$$

$$\Rightarrow \quad \sigma_k^2 = \frac{1}{N_k} \sum_{i \in D} r_{ik} \big(x^{(i)} - \mu_k\big)^2 \tag{17}$$

Obtaining optimal π_k:

Minimizing Equation 15 w.r.t π_k is equivalent to maximizing $\sum_{i \in D} log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)$ w.r.t π_k. Here, we must take account of the constraint $\sum_{k=1}^{K} \pi_k = 1$, which requires the mixing coefficients to sum to one. To achieve this, Lagrange multiplier is introduced:

$$\sum_{i \in D} log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k) + \lambda \Big(\sum_{k=1}^{K} \pi_k - 1 \Big) \tag{18}$$

The above equation has to be maximized to obtain optimal π_k.

$$\frac{\partial}{\partial \pi_k} : \sum_{i \in D} \frac{\mathcal{N}(x^{(i)}|\mu_k, \sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'}, \sigma_{k'})} + \lambda = 0 \tag{19}$$

Multiplying the above equation by π_k and summing over k and using $\sum_{k=1}^{K} \pi_k = 1$, we obtain $\lambda = -N$ where N is total number of samples in the dataset D. Now, multiplying Equation 19 by π_k and replacing λ with $-N$, we get

$$\sum_{i \in D} \pi_k \frac{\mathcal{N}(x^{(i)}|\mu_k, \sigma_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x^{(i)}|\mu_{k'}, \sigma_{k'})} - \pi_k N = 0$$

$$\Rightarrow \quad \sum_{i \in D} r_{ik} - \pi_k N = 0 \qquad (\textit{Using Equation 14})$$

$$\Rightarrow \quad \pi_k = \frac{N_k}{N} \tag{20}$$

Equation 16, 17, 20 gives the optimal value of μ_k, σ_k and π_k respectively, however these equations assume that posterior on latent variable i.e. r_{ik} is known which is not true. Hence, it's not a closed form solution as we need an iterative algorithm to estimate r_{ik} and accordingly update μ_k, σ_k and π_k over several iterations.

4.3 Expectation Maximization

The iterative scheme for finding a solution to the minimizing log-likelihood problem turns out to be an instance of the EM algorithm for the particular case of the Gaussian mixture model. First we choose some initial values for the means, covariances, and mixing coefficients. Then we alternate between the following two updates, called the E step and M step. In the expectation step, or E step, we use the current values for the parameters to evaluate the posterior probabilities, or responsibilities, given by Equation 14. We then use these probabilities in the maximization step, or M step, to re-estimate the means, covariances, and mixing coefficients using the results 16, 17, and 20. Formally, the algorithm is defined below:

Algorithm: EM for Gaussian mixture model

---

1. **Input**: Dataset consisting of $x^{(i)}$ samples with $i = 1, 2, ..., N$
2. **Initialization**: Initialize μ, σ and π
3. Iteration Loop:
   E step: Evaluate posteriors using current parameter estimates

$$r_{ik} = \frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^{(i)}|\mu_{\hat{k}}, \sigma_{\hat{k}})}$$

   M step: Update the parameters using the posterior obtained from E-step

$$\mu_k = \frac{1}{N_k} \sum_{i \in D} r_{ik} x^{(i)}$$

$$\sigma_k^2 = \frac{1}{N_k} \sum_{i \in D} r_{ik} \left(x^{(i)} - \mu_k\right)^2$$

$$\pi_k = \frac{N_k}{N}$$

Description of figure 2: It shows EM algorithm for mixture of Gaussians for two Gaussian components. Plot (a) shows the data points in green, together with the initial configuration for the two Gaussian components are shown as blue and red circles. Plot (b) shows the result of the initial E step, in which each data point is depicted using a proportion of blue ink equal to the posterior probability of having been generated from the blue component, and a corresponding proportion of red ink given by the posterior probability of having been generated by the red component. Thus, points that have a significant probability for belonging to either cluster appear purple. The situation after the first M step is shown in plot (c), in which the mean of the blue Gaussian has moved to the mean of the data set, weighted by the probabilities of each data point belonging to the blue cluster, in other words it has moved to the centre of mass of the blue ink. Similarly, the covariance of the blue Gaussian is set equal to the covariance of the blue ink. Analogous results hold for the red component. Plots (d), (e), and (f) show the results after 2, 5, and 20 iterations of EM, respectively. In plot (f) the algorithm is close to convergence.

4.4 Similarity with K-Means Clustering

In the last lecture, we studied about k-Means, a simple iterative algorithm to divide the given data points into a group of K clusters, for a fixed $K \in N$. Following are few similarities of the Gaussian Mixture model with the K-means clustering.

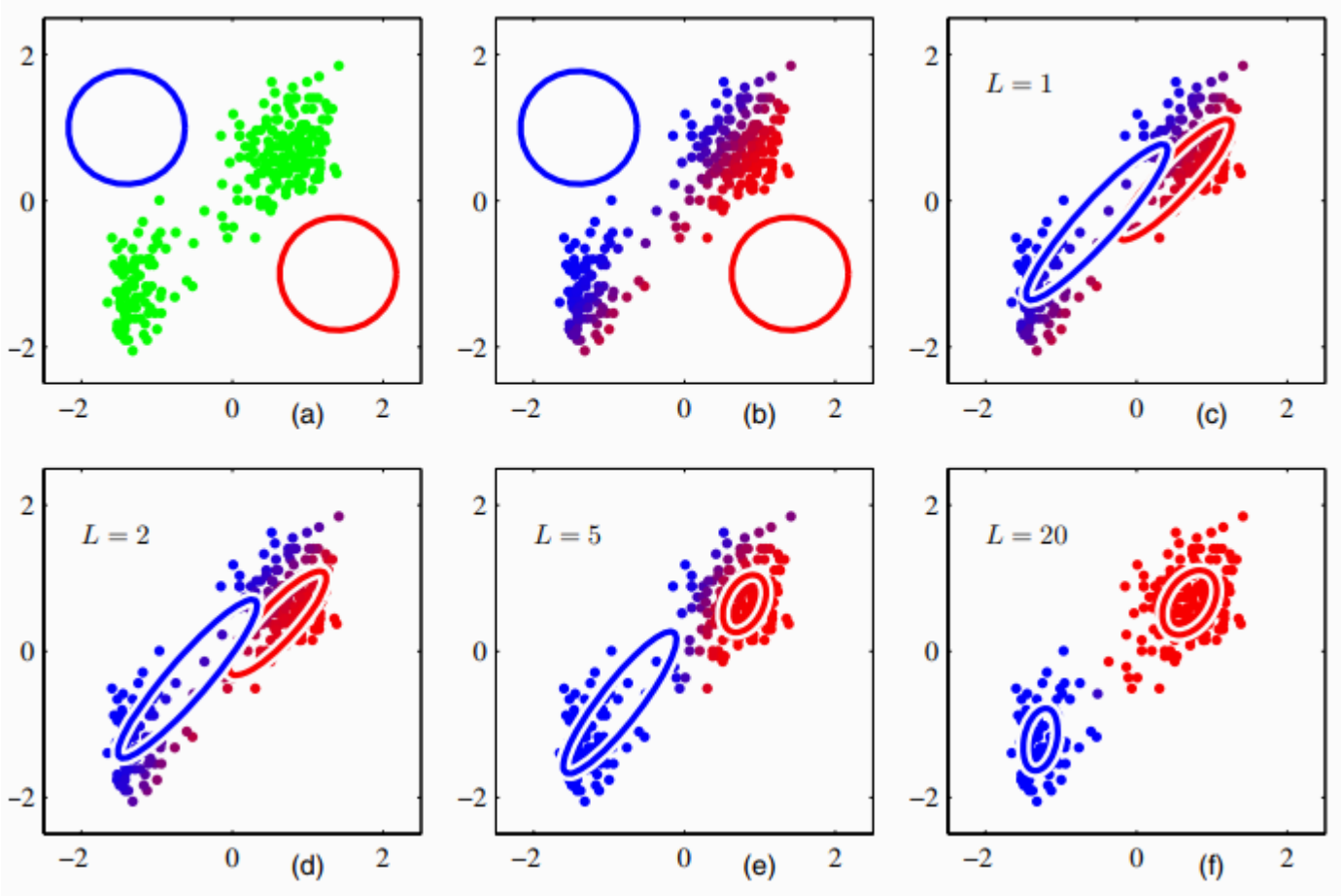

Figure 2: Illustration of EM algorithm across iterations

- Posterior for the latent variable, r_{ik} is an assignment of sample i to cluster k. Since it takes values in the range of [0, 1], it's a soft assignment unlike the case for K-means where assignment is {0, 1} i.e. hard assignment.
- Model parameters, π_k are cluster centers for kth cluster.

4.4.1 Making the similarity formal

If we fix $\sigma_k^2 = \epsilon \ \forall \ k$ in the Gaussian mixture set-up, Equation 14 becomes:

$$r_{ik} = \frac{\pi_k \exp\Big(-\frac{1}{2\epsilon}(x^{(i)} - \mu_k)^2\Big)}{\sum_{k'=1}^{K} \pi'_k \exp\Big(-\frac{1}{2\epsilon}(x^{(i)} - \mu'_k)^2\Big)} \tag{21}$$

If we consider the limit $\epsilon \to 0$, we see that in the denominator the term for which $(x^{(i)} - \mu'_k)^2$ is smallest will go to zero most slowly, and hence the posteriors or responsiblities r_{ik} for the data point $x^{(i)}$ all go to zero except for term k', for which the posterior, $r_{ik'}$ will go to unity. Note that this holds independently of the values of the π_k so long as none of the π_k is zero. Thus, in this limit, we obtain a hard assignment of data points to clusters, just as in the K-means algorithm. Each data point is thereby assigned to the cluster having the closest mean. The EM re-estimation

for the μ_k term simply reduces to the K-means result.
In the limit $\epsilon \to 0$, the cost function (to be minimized) becomes

$$\sum_{i \in D} \sum_{k=1}^{K} \frac{1}{2} r_{ik} \left(x^{(i)} - \mu_k\right)^2 \tag{22}$$

which is same as that of K-means algorithm.

5 Summary

This lecture introduces generative modeling and discusses about the differences between generative and discriminative paradigms of statistical learning. It talks in detail about a generative modeling technique, called Gaussian Mixture Model (GMM). It describes how introduction of latent variables make maximizing the likelihood for given data more tractable. Further, it shows how Expectation-Maximization (EM) algorithm are used for learning their parameters. Finally, it discusses about similarity between the GMM and K-means clustering and how GMM can be reduced to K-means under certain conditions.

References

[1] N. M. Nasrabadi. Pattern recognition and machine learning. *Journal of electronic imaging*, 16(4):049901, 2007.