University of Pittsburgh

ECE 2195: Special Topics – Computers Machine Learning

Mixture of Gaussian Models

Mai Abdelhakim, PhD

ECE Department

Swanson School of Engineering

University of Pittsburgh

maia@pitt.edu

Recall - ML estimates of parameters of classes can be obtained from data

- Taking the $\log_e$ of the likelihood function & adding constraints that probability sum to 1
- The derivative w.r.t (with respect to) each of the parameters (priors, means, variances)

- We get
  - $\Pi_k = \frac{n_k}{\sum_{k=1}^{K} n_k} = \frac{n_k}{n}$
    - n_k is the number of samples from class k
  - $\mu_{k,f} = \frac{\sum_{i:y_i=k} x_{i,f}}{n_k}$, mean of class k feature f
  - $\sigma_{k,f}^2 = \frac{\sum_{i:y_i=k}(x_{i,f} - \mu_{k,d})^2}{n_k}$, variance of class k feature f

The total no. of samples: n= $\sum_{k=1}^{K} n_k$

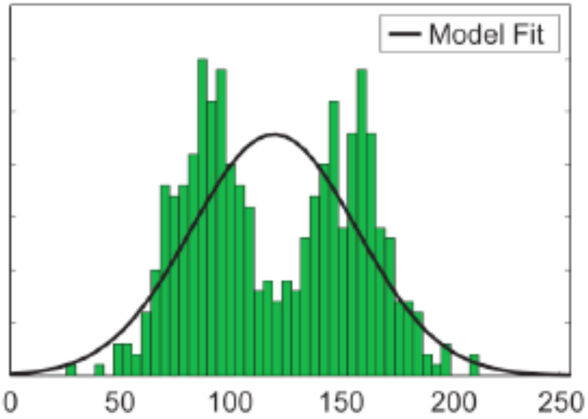
What if observations are not labeled?

Let the number of classes is known to be K!

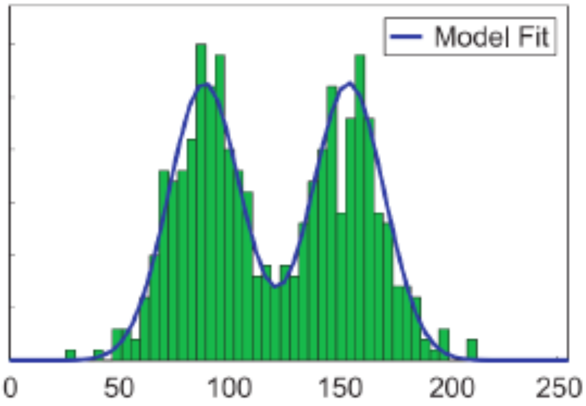
Let's also assume that features in each class can be modeled as Gaussian

Mixture of Gaussian

2 different classes



Fitting data into one Gaussian model



Fitting data into a mixture of Gaussian models

Ref: K. Kutulakos

Unknown Class Labels – Use Latent Variable

- The Gaussian models or observations are **not labeled** no y_i

- For each observation x define latent variable z vector of length K with kth element $\boldsymbol{z_k}$ =1 if observation should belong to class k

  - $\boldsymbol{z_k} = \mathbf{1}$ is a flag that observation is from **class k**. (it is =1 for only one k and zero for the rest)

$$p(z_k = 1) = \pi_k$$

$z_k = 1$ if class is k
similar to responsibility

$$0 \leq \pi_k \leq 1,\ \sum_{k=1}^{K} \pi_k$$

Mixture of Gaussian Models

- The conditional probability of a feature vector x in each class is Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

one class

- But classes are not labeled – we get mixture of Gaussians

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Responsibility is the posterior probability

- The posterior probability (also called responsibility)

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

After observing feature x what is
the probability that is from class k

unlabeled

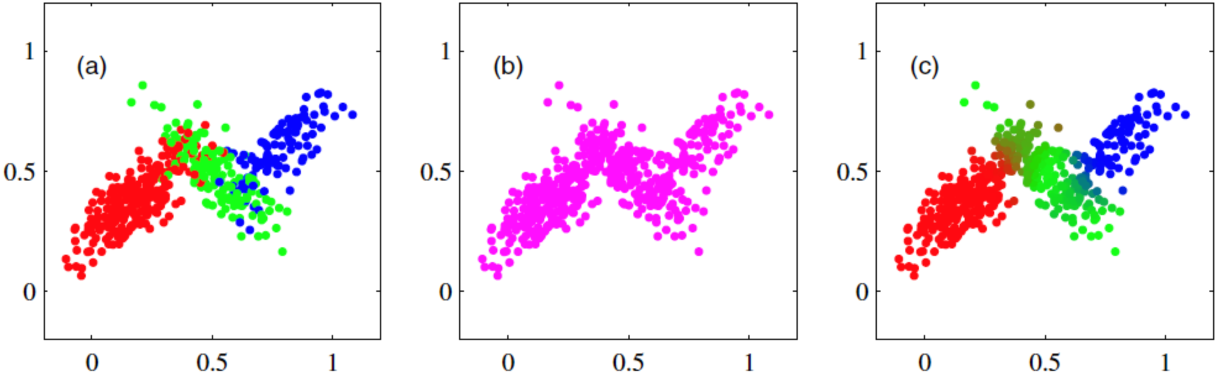


Figure 9.5 Example of 500 points drawn from the mixture of 3 Gaussians shown in Figure 2.23. (a) Samples from the joint distribution $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ in which the three states of $\mathbf{z}$, corresponding to the three components of the mixture, are depicted in red, green, and blue, and (b) the corresponding samples from the marginal distribution $p(\mathbf{x})$, which is obtained by simply ignoring the values of $\mathbf{z}$ and just plotting the $\mathbf{x}$ values. The data set in (a) is said to be *complete*, whereas that in (b) is *incomplete*. (c) The same samples in which the colours represent the value of the responsibilities $\gamma(z_{nk})$ associated with data point $\mathbf{x}_n$, obtained by plotting the corresponding point using proportions of red, blue, and green ink given by $\gamma(z_{nk})$ for $k = 1, 2, 3$, respectively

Reference: Bishop, chapter 9

The Likelihood Function - With the i.i.d assumption

- Recall that for an observation x

K classes

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\ln\left(\prod_{n=1}^{N} p(x_n)\right) = \sum_{n=1}^{N} \ln(p(x_n))$$

- For all observation X, $P(X) = \prod_{n=1}^{N} p(x_n)$. (N is all the samples or observations) – The maximum likelihood can be obtained, as:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \overbrace{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}^{p(x_n)} \right\}$$

$$\frac{\partial}{\partial \mu_k} = 0$$

Taking the ln will not cancel the exponent in the Gaussian due to the mixture

How to optimize this? Every observation has a latent variable z

Taking the derivative w.r.t. mean μ_k

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

N = total Number of samples

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

$\boldsymbol{z_{nk}} = \mathbf{1}$ is a flag that observation n is from **class k**.

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \underbrace{\gamma(z_{nk})}_{\text{probability that it lies in class k}} \mathbf{x}_n$$

probability that it lies in class k

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$Note: \quad \frac{\partial(\mathbf{x}^T A \mathbf{x})}{\partial \mathbf{x}} = \mathbf{x}^T(A + A^T)$$

Defining the objective and log likelihood under constraint that total probability of classes is 1, we find derivative and get

- For every k, we get

N = total Number of samples

Know number of classes

$$\boxed{\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n}$$

$$\boxed{\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\mu_k)(\mathbf{x}_n-\mu_k)^{\mathrm{T}}}$$

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

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

Expectation Maximization

- No closed form solution for the problem - the responsibility depends on the parameters

- Need an iterative solution – like **Expectation Maximization technique**

EM – Init and E-step

Randomly

- Initialize parameters : means, covariances and priors of all classes

- E step: (find responsibilities)
  - Which Gaussian generated each datapoint?
  - It's a distribution over all possibilities

$$
\begin{aligned}
\gamma_k = p(z=k|\mathbf{x}) &= \frac{p(z=k)p(\mathbf{x}|z=k)}{p(\mathbf{x})} \\
&= \frac{p(z=k)p(\mathbf{x}|z=k)}{\sum_{j=1}^{K} p(z=j)p(\mathbf{x}|z=j)} \\
&= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j,\Sigma_j)}
\end{aligned}
$$

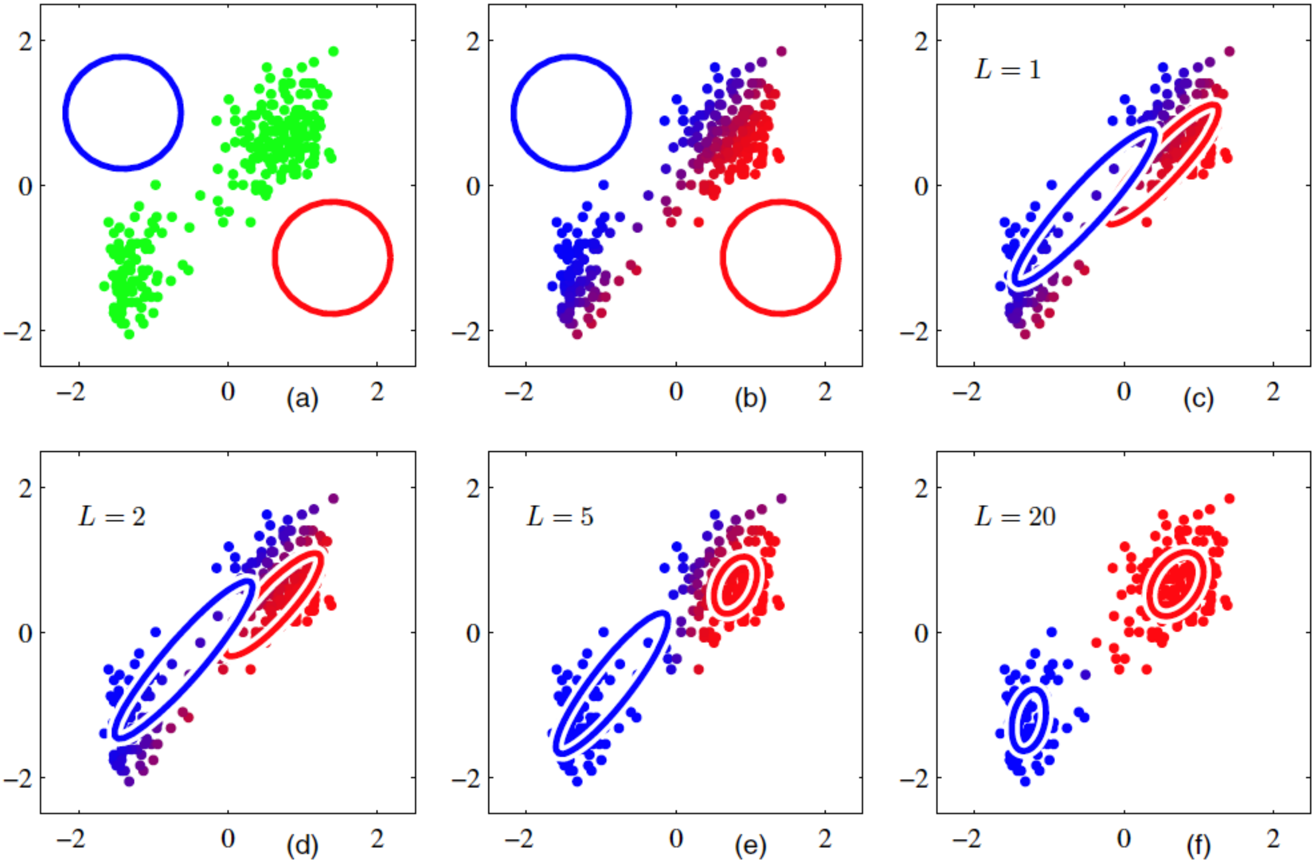
- M step: re-estimate parameters
  - Optimal point has zero gradient

$$
\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \\
\pi_k^{\text{new}} &= \frac{N_k}{N}
\end{aligned}
$$

EM – Convergence

- Check converges by checking log likelihood (or the parameters' values stop changing)

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}^{(n)}|\mu_k, \Sigma_k) \right)$$



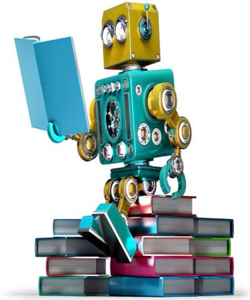
2
0
−2
−2
0
2
(a)
(b)
(c)
(d)
(e)
(f)
$L = 1$
$L = 2$
$L = 5$
$L = 20$

Learning from Data

- It turns out that machines can do a lot!
  - Enable automation
  - Less expensive

- Learning depends on Data
  - With Internet of Things (IoT), massive amount of data can be collected

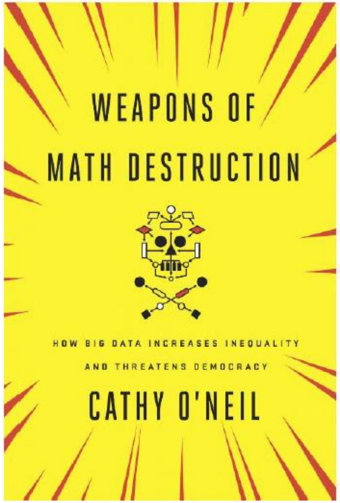
- Bad data → bad model!



Ethical Considerations

- Preserve privacy
  - How to handle sensitive data
    - Can we identify people and/or their location from data set?
  - Privacy preserving techniques (Example: K – anonymity)

- Avoid creating biased models : create inequalities
  - Data collection can be biased:
    - Example – Hiring decisions by machine learning: if data from particular gender or ethnicity group is dominant in a dataset, this may affect hiring decisions
  - Avoid training models that repeat mistakes happened in the past
  - Until now, no policies govern these issues
  - **Book: “Weapons of Math Destruction” by Cathy O'Neil**



TED Talk Video: https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end/transcript