Latent Variables, Mixture Models and EM

Christopher M. Bishop
Microsoft Research, Cambridge

BCS Summer School
Exeter, 2003
Overview.

- K-means clustering
- Gaussian mixtures
- Maximum likelihood and EM
- Latent variables: EM revisited
- Bayesian Mixtures of Gaussians
Old Faithful
Old Faithful Data Set

Time between eruptions (minutes)

Duration of eruption (minutes)
K-means Algorithm

• Goal: represent a data set in terms of K clusters each of which is summarized by a prototype $\mu_k$

• Initialize prototypes, then iterate between two phases:
  – E-step: assign each data point to nearest prototype
  – M-step: update prototypes to be the cluster means

• Simplest version is based on Euclidean distance
  – re-scale Old Faithful data
Responsibilities

• *Responsibilities* assign data points to clusters

\[ r_{nk} \in \{0, 1\} \]

such that

\[ \sum_k r_{nk} = 1 \]

• Example: 5 data points and 3 clusters

\[
(r_{nk}) = \begin{pmatrix}
1 & 0 & 0 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
1 & 0 & 0
\end{pmatrix}
\]
K-means Cost Function

\[
J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|x_n - \mu_k\|^2
\]
Minimizing the Cost Function

- **E-step**: minimize $J$ w.r.t. $r_{nk}$
  - assigns each data point to nearest prototype
- **M-step**: minimize $J$ w.r.t $\mu_k$
  - gives
    \[
    \mu_k = \frac{\sum_n r_{kn} x_n}{\sum_n r_{kn}}
    \]
  - each prototype set to the mean of points in that cluster
- Convergence guaranteed since there is a finite number of possible settings for the responsibilities
Limitations of K-means

- Hard assignments of data points to clusters – small shift of a data point can flip it to a different cluster
- Not clear how to choose the value of K
- Solution: replace ‘hard’ clustering of K-means with ‘soft’ probabilistic assignments
- Represents the probability distribution of the data as a Gaussian mixture model
The Gaussian Distribution

• Multivariate Gaussian

\[ \mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi|\Sigma|)^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\} \]

mean

covariance

• Define precision to be the inverse of the covariance

\[ \Lambda = \Sigma^{-1} \]

• In 1-dimension

\[ \tau = \frac{1}{\sigma^2} \]
Likelihood Function

- Data set

\[ D = \{x_n\} \quad n = 1, \ldots, N \]

- Assume observed data points generated independently

\[
p(D | \mu, \Sigma) = \prod_{n=1}^{N} \mathcal{N}(x_n | \mu, \Sigma)
\]

- Viewed as a function of the parameters, this is known as the \textit{likelihood function}
Maximum Likelihood

• Set the parameters by maximizing the likelihood function
• Equivalently maximize the log likelihood

\[
\ln p(D | \mu, \Sigma) = -\frac{N}{2} \ln |\Sigma| - \frac{N}{2} \ln(2\pi) \\
-\frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu)
\]
Maximum Likelihood Solution

- Maximizing w.r.t. the mean gives the *sample mean*

\[ \mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

- Maximizing w.r.t covariance gives the *sample covariance*

\[ \Sigma_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\text{ML}})(x_n - \mu_{\text{ML}})^\top \]
Gaussian Mixtures

- Linear super-position of Gaussians
  
  \[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \]

- Normalization and positivity require
  
  \[ \sum_{k=1}^{K} \pi_k = 1 \quad 0 \leq \pi_k \leq 1 \]

- Can interpret the mixing coefficients as prior probabilities
  
  \[ p(x) = \sum_{k=1}^{K} p(k)p(x|k) \]
Example: Mixture of 3 Gaussians
Contours of Probability Distribution
Sampling from the Gaussian

- To generate a data point:
  - first pick one of the components with probability $\pi_k$
  - then draw a sample $x_n$ from that component
- Repeat these two steps for each new data point
Synthetic Data Set
Fitting the Gaussian Mixture

• We wish to invert this process – given the data set, find the corresponding parameters:
  – mixing coefficients
  – means
  – covariances

• If we knew which component generated each data point, the maximum likelihood solution would involve fitting each component to the corresponding cluster

• Problem: the data set is unlabelled

• We shall refer to the labels as *latent* (= hidden) variables
Synthetic Data Set Without Labels
Posterior Probabilities

• We can think of the mixing coefficients as prior probabilities for the components
• For a given value of $x$ we can evaluate the corresponding posterior probabilities, called responsibilities
• These are given from Bayes’ theorem by

$$
\gamma_k(x) \equiv p(k|x) = \frac{p(k)p(x|k)}{p(x)} = \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}
$$
Posterior Probabilities (colour coded)
Posterior Probability Map
Maximum Likelihood for the GMM

- The log likelihood function takes the form

\[
\ln p(D|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right\}
\]

- Note: sum over components appears \textit{inside} the log
- There is no closed form solution for maximum likelihood
Problems and Solutions

• How to maximize the log likelihood
  – solved by expectation-maximization (EM) algorithm
• How to avoid singularities in the likelihood function
  – solved by a Bayesian treatment
• How to choose number K of components
  – also solved by a Bayesian treatment
EM Algorithm – Informal Derivation

- Let us proceed by simply differentiating the log likelihood
- Setting derivative with respect to $\mu_j$ equal to zero gives

$$- \sum_{n=1}^{N} \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j) \left( \frac{\sum_k \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\gamma_j(x_n)} \right) \Sigma_j^{-1} (x_n - \mu_j) = 0$$

giving

$$\mu_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n)x_n}{\sum_{n=1}^{N} \gamma_j(x_n)}$$

which is simply the weighted mean of the data
EM Algorithm – Informal Derivation

• Similarly for the covariances

\[ \Sigma_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n)(x_n - \mu_j)(x_n - \mu_j)^T}{\sum_{n=1}^{N} \gamma_j(x_n)} \]

• For mixing coefficients use a Lagrange multiplier to give

\[ \pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(x_n) \]
EM Algorithm – Informal Derivation

• The solutions are not closed form since they are coupled

• Suggests an iterative scheme for solving them:
  – Make initial guesses for the parameters
  – Alternate between the following two stages:
    1. E-step: evaluate responsibilities
    2. M-step: update parameters using ML results
EM – Latent Variable Viewpoint

• Binary latent variables $z = \{z_{kn}\}$ describing which component generated each data point

• Conditional distribution of observed variable

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Lambda_k)^{z_k}$$

• Prior distribution of latent variables

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$

• Marginalizing over the latent variables we obtain

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Lambda_k)$$
Expected Value of Latent Variable

- From Bayes’ theorem

\[
E[z_{ni}] = \frac{\sum_{\mathcal{Z}_ni} z_{ni} [\pi_ip(x_n|\mu_i, \Sigma_i)]z_{ni}}{\sum_{\mathcal{Z}_ni} [\pi_ip(x_n|\mu_i, \Sigma_i)]z_{ni}}
\]

\[
= \frac{\pi_i \mathcal{N}(x|\mu_i, \Sigma_i)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}
\]

\[
= \gamma_i(x_n)
\]
Complete and Incomplete Data

complete

incomplete
Latent Variable View of EM

• If we knew the values for the latent variables, we would maximize the complete-data log likelihood

\[ \ln p(x, z|\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{ \ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k) \} \]

which gives a trivial closed-form solution (fit each component to the corresponding set of data points)

• We don’t know the values of the latent variables

• However, for given parameter values we can compute the expected values of the latent variables
Expected Complete-Data Log Likelihood

- Suppose we make a guess $\theta_{\text{old}}$ for the parameter values (means, covariances and mixing coefficients)
- Use these to evaluate the responsibilities
- Consider expected complete-data log likelihood

$$\mathbb{E}_z[\ln p(x, z|\theta)] = \sum_{n=1}^{N} \sum_{i=1}^{K} \gamma_i(x_n) \{ \ln \pi_i + \ln \mathcal{N}(x_n|\mu_i, \Sigma_i) \}$$

where responsibilities are computed using $\theta_{\text{old}}$
- We are implicitly ‘filling in’ latent variables with best guess
- Keeping the responsibilities fixed and maximizing with respect to the parameters give the previous results
EM in General

- Consider arbitrary distribution $q(z)$ over the latent variables.
- The following decomposition always holds:

$$\ln p(x|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q||p)$$

where

$$\mathcal{L}(q, \theta) = \sum_z q(z) \ln \left\{ \frac{p(x, z|\theta)}{q(z)} \right\}$$

$$\text{KL}(q||p) = -\sum_z q(z) \ln \left\{ \frac{p(z|x, \theta)}{q(z)} \right\}$$
Decomposition

\[ KL(q||p) \]

\[ L(q,\theta) \]

\[ \ln p(x|\theta) \]
Optimizing the Bound

- E-step: maximize $\mathcal{L}$ with respect to $q(z)$
  - equivalent to minimizing KL divergence
  - sets $q(z)$ equal to the posterior distribution $p(z|x, \theta)$
- M-step: maximize bound with respect to $\theta$
  - equivalent to maximizing expected complete-data log likelihood
- Each EM cycle must increase incomplete-data likelihood unless already at a (local) maximum
E-step

\[ \text{KL}(q \| p) = 0 \]

\[ \mathcal{L}(q, \theta_{\text{old}}) \]

\[ \ln p(x \| \theta_{\text{old}}) \]
M-step

\[ \text{KL}(q \| p) \]

\[ \mathcal{L}(q, \theta_{\text{new}}) \]

\[ \ln p(x | \theta_{\text{new}}) \]