Statistical and Learning Techniques in Computer Vision Lecture 4: Gaussian Mixture Models and the EM Algorithm

Jens Rittscher and Chuck Stewart

1 Motivation

- We will continue with our problem of modeling densities.
- Similar to kernel density estimates, we will form the density function from a linear combination of basis densities, where the basis densities are written parametrically. Here the number of basis functions, K, will be much less than the number of given data points.
- Most often in using mixture models there is a significance to each basis functions — such as when points are sampled from a set of lines.
- We use the mixture model to introduce the Expectation-Maximization (EM) algorithm.

2 Mixture models

- A mixture model is a linear combination of K densities:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\theta}_k) \tag{1}$$

  where

  - The k values are discrete labels of the individual densities. We can also think of the k values as class labels.
  - π_k is the *discrete* probability that a point is sampled from class k. The π_k values are also thought of as the *mixture parameters*.
  - $p(\mathbf{x}|\boldsymbol{\theta}_k)$ are the *component densities* — the probability that $\mathbf{x}$ takes on certain values given that it is from class k. They are also called *class conditional densities*.

- Here are the constraints on the probabilities:

$$\sum_{k=1}^{K} \pi_k = 1 \qquad \text{with} \qquad 0 \leq \pi_k \leq 1, \;\; \text{for } k = 1, \ldots, K \tag{2}$$

  and

$$\int p(\mathbf{x}|\boldsymbol{\theta}_k) dx = 1, \;\; \text{for } k = 1, \ldots, K \tag{3}$$

- To help clarify the structure of the mixture model, here is how a set of points would be generated (sampled) from a mixture model with specific values for π_k and $\boldsymbol{\theta}_k$ fixed for all k:
  - Randomly select the component $\hat{k}$ from which to sample using the probabilities $\{\pi_1, \ldots, \pi_K\}$
  - Using the parameters $\boldsymbol{\theta}_{\hat{k}}$ and the form of the density associated with the selected $\hat{k}$, generate the data point $\mathbf{x}$ according to

$$p(\mathbf{x}|\boldsymbol{\theta}_{\hat{k}}) \tag{4}$$

3 Our Problem

- The problem we are going to focus on here is estimating the mixture parameters

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K) \tag{5}$$

  and the parameters of all the individual densities

$$\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K). \tag{6}$$

  In the case of a mixture of Gaussians (normal distributions) we will write $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in place of $\boldsymbol{\theta}$, where

$$\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) \tag{7}$$

  and

$$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k) \tag{8}$$

  are the mean vectors and covariance matrices of the component (multivariate) normal distributions.

- We are given a set of N data vectors $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Each of these is assumed to be drawn independently from the underlying mixture distribution that we are trying to model.

- What we do not know is the assignment of the data vectors $\mathbf{x}_n$ to the individual components, $k = 1, \ldots, K$. If we did estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ would be a simple exercise in maximum likelihood estimation, applied once for each k.

- Without this knowledge, we can write down a log-likelihood function:

$$\log p(\mathcal{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9}$$

  We will see that we can use the EM-algorithm so solve this maximum likelihood estimation problem.

4 Modeling Using Hidden Variables

- We introduce *hidden variables* to represent the unknown assignment of data points to models.
  - These are also called *latent variables*.
- For particular $\mathbf{x}$, we define a k-dimensional **binary** vector, $\mathbf{z}$, which has the following properties:
  - $z_k \in \{0, 1\}$, $k = 1, \ldots, K$.
  - $z_k = 1$ represents the assignment of point $\mathbf{x}$ to component k.
  - If $z_k = 1$ then $z_j = 0$ for $j \neq k$.
- At first, this representation seems wasteful, since only one component of $\mathbf{z}$ is non-zero. But, during estimation of the parameters, for each data point we will want to model the probability that point n is from component k. This is the probability that $z_{n,k} = 1$. These probabilities will be generally be non-zero for each k. We will discuss this further shortly.
- To understand the meaning of $\mathbf{z}$ a bit further and to prepare for the derivation of EM, here are some additional properties. Many of these follow from the binary nature of $\mathbf{z}$, including the fact that $\mathbf{z}_k = 1$ for exactly one value of k:
  - $p(z_k = 1|\boldsymbol{\pi}) = \pi_k$. This can be thought of as the prior probability (if $\boldsymbol{\pi}$ is known) that a particular point comes from component k.
  - We can write
    $$p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} {\pi_k}^{z_k}. \tag{10}$$
    Note that this does not depend on the normal distribution parameters $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$.
  - Lastly, we have the conditional probability
    $$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{k=1}^{K} (\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{n,k}}. \tag{11}$$
    Note that $\boldsymbol{\pi}$ does not play a role on the right-hand side because $\mathbf{z}$ is fixed.
- Using these properties, we can write

$$\begin{aligned} p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\mathbf{z}|\boldsymbol{\pi}) \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k \boldsymbol{\Sigma}_k). \end{aligned} \tag{12}$$

- Finally, consider the posterior probability that $z_{n,k} = 1$ for particular data point n and component k. This is so important that we will use a special function $\gamma(z_{n,k})$ to represent it. Using Bayes' rule we have

$$\begin{aligned}\gamma(z_{n,k}) &\equiv p(z_{n,k} = 1|\mathbf{x}_n, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{p(z_{n,k} = 1|\boldsymbol{\pi})p(\mathbf{x}_n|z_{n,k} = 1, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sum_{j=1}^{K} p(z_{n,j} = 1|\boldsymbol{\pi})p(\mathbf{x}_n|z_{n,j} = 1, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}\end{aligned} \tag{13}$$

5 Maximum Likelihood

- We will approach the problem of estimating $\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\Sigma}$ from a maximum likelihood perspective.
- Before doing so, we consider an important problem with the likelihood function we have written in (9): It has singularities, where its value can go to infinity!
  - This occurs when $\boldsymbol{\mu}_k \rightarrow \mathbf{x}_n$, for some k and n, and $\boldsymbol{\Sigma}_k = \sigma^2 \mathbf{I}$ with $\sigma^2 \rightarrow 0$.

  We will discuss this in class and show why the same problem does not occur with maximum likelihood estimation for single normal distributions.
- The solution is to place a prior on the values of $\boldsymbol{\Sigma}_k$ and proceed with MAP estimation.
- Our discussion will proceed here ignoring the problem.

6 EM for Gaussian Mixtures — Informal Derivation

- Using the likelihood expression from (9), repeated here for clarity,

$$\log p(\mathcal{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{14}$$

  we will derive conditions on the maximum of the likelihood with respect to $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.
- Taking the derivative with respect to $\boldsymbol{\mu}_k$ for each k, setting result equal to $\mathbf{0}$ and solving produces the condition,

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma(z_{n,k})\mathbf{x}_n}{N_k}, \tag{15}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{n,k}) = \tag{16}$$

is the *expected number of points assigned to component* k.

- Taking the derivative with respect to $\mathbf{\Sigma}_k$ for each k, setting the result to the $d \times d$ 0 matrix, and solving produces the condition,

$$\mathbf{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma(z_{n,k})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top}{N_k}, \tag{17}$$

- Finally, imposing the constraint $\sum_{k=1}^{K} \pi_k = 1$ using Lagrange multipliers, we can also derive the condition

$$\boldsymbol{\pi}_k = \frac{N_k}{N} \tag{18}$$

- Comments:
  - We will show the derivation of these in class.
  - Throughout we exploit a number of independence assumptions.
  - Notice how the posterior probabilities of $z_{n,k}$ have snuck in here, even though the hidden variables $\mathbf{z}$ are not part of the likelihood formulation.

- The results are remarkably simple except for the following crucial fact:
  - The $\gamma(z_{n,k})$ values depend on $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ and the values of $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ depend on the $\gamma(z_{n,k})$.

- We address this problem using iterative update equations, which is the expectation maximization (EM) algorithm for Gaussian mixture models.
  1. Compute initial estimates for $\boldsymbol{\pi}^0$, $\boldsymbol{\mu}^0$ and $\mathbf{\Sigma}^0$. Set $t = 0$
  2. **E step:** Using $\boldsymbol{\pi}^t$, $\boldsymbol{\mu}^t$ and $\mathbf{\Sigma}^t$, compute new values $\gamma^{t+1}(z_{n,k})$ for $n = 1, \ldots, N$ and $k = 1, \ldots, K$ as specified in (13).
  3. **M step:** Using the resulting $\gamma^{t+1}(z_{n,k})$ values, compute new estimates $\boldsymbol{\pi}^{t+1}$, $\boldsymbol{\mu}^{t+1}$, and $\mathbf{\Sigma}^{t+1}$ as specified in (15), (17) and (18).
  4. Check for convergence of the resulting log likelihood function (9) or a "sufficiently small" change in the estimated parameters. If convergence is reached, quit. Otherwise, set $t = t + 1$, and repeat steps 2-4.

7 Initialization

- Initialization is difficult in general. This includes both selecting K, the number of components, and the initial values of $\boldsymbol{\pi}^0$, $\boldsymbol{\mu}^0$ and $\mathbf{\Sigma}^0$.

- Here we will use the K-means algorithm, described as follows.

1. The starting point is K seed locations, $\mathbf{y}_1, \ldots, \mathbf{y}_K$.
2. For each $\mathbf{x}_n$, find the $\mathbf{y}_k$ it is closest to. This determines an initial value of the binary vector $\mathbf{z}_n$ — i.e. which one of the K entries in $\mathbf{z}_n$ is assigned a 1 instead of a 0.
3. For each k, compute
$$\mathbf{y}_k = \frac{\sum_{n=1}^{N} z_{n,k} \mathbf{x}_n}{\sum_{n=1}^{N} z_{n,k}}. \tag{19}$$
More simply, this is the mean of the points assigned to the k-th component.
4. Repeat the assignment of points to components (the computation of $z_{n,k}$) and the recalculation of $\mathbf{y}_k$, until convergence.

- Using the results, the initialization of $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is straight-forward. In particular,
$$\pi_k^0 = \frac{\sum_{n=1}^{N} z_{n,k}}{N}, \tag{20}$$
which is just the fraction of points assigned to the k-th mean,
$$\boldsymbol{\mu}_k^0 = \mathbf{y}_k \tag{21}$$
and
$$\boldsymbol{\Sigma}_k^0 = \frac{\sum_{n=1}^{N} z_{n,k} (\mathbf{x}_n - \boldsymbol{\mu}_k^0)(\mathbf{x}_n - \boldsymbol{\mu}_k^0)^\top}{\sum_{n=1}^{N} z_{n,k}}. \tag{22}$$
These are the maximum likelihood estimates computed from the points assigned to the k component. Their close relationship to the EM update equations should be clear.

8 More Theory

- Re-examine the two types of data we have:
  - $\mathcal{X}$ is the known set of data measurements
  - $\mathcal{Z}$ is the set of *unknown* assignments of data points to model components.
- $\boldsymbol{\Theta}$ is our vector of unknown parameters.
- The log likelihood we need to maximize is
$$\log p(\mathcal{X}|\boldsymbol{\Theta}) = \log \sum_{\mathcal{Z}} p(\mathcal{X}, \mathcal{Z}|\boldsymbol{\Theta}) \tag{23}$$
  - Note that if we knew $\mathcal{Z}$, then maximizing $p(\mathcal{X}, \mathcal{Z}|\boldsymbol{\Theta})$ would be easy.
  - Without knowing $\mathcal{Z}$, it is generally quite hard to evaluate the log of sum taken over all of the possible $\mathcal{Z}$.

- The trick of EM is to manipulate this log-sum combination into the form

  $$\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Theta}^{\text{old}}) = \sum_{\mathcal{Z}} p(\mathcal{Z}|\mathcal{X}, \mathbf{\Theta}^{\text{old}}) \log p(\mathcal{X}, \mathcal{Z}|\mathbf{\Theta}) \tag{24}$$

  where

  - $\mathbf{\Theta}^{\text{old}}$ is a previous estimate of the parameters
  - $p(\mathcal{Z}|\mathcal{X}, \mathbf{\Theta}^{\text{old}})$ is the posterior probability of the hidden variables of $\mathcal{Z}$ given the observed data and the previous estimate of the parameters.
  - The summation computes the *expected value of the log-likelihood* of $\mathcal{X}$ and $\mathcal{Z}$ given the posterior probability of the hidden variables. This sum is easier to evaluate than the sum of the logs.

- The evaluation works in two steps:

  - Computation of the expected value function $\mathcal{Q}$ given a particular $\mathbf{\Theta}^{\text{old}}$. This the expectation or “E” step.
  - Maximization of $\mathcal{Q}$ with respect to $\mathbf{\Theta}$. This is the “M” step.

- In practice, the computation often reduces to the computation of the $\gamma(z_{n,k})$ values, which are the *expected values* of the hidden variables given the parameters and the data, followed by re-estimation of the parameters.

  - This is exactly what we did above for Gaussian mixture models.

- Convergence (to local maxima) can be provided in several ways, one involving the use of Jensen’s inequality, and the other involving approximation of distributions and comparison of distributions through the Kullback-Liebler divergence.

9 Further Reading

Most of this lecture is adapted from [Bis06, Chapter 9]

References

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.