Introduction to Statistical Machine Learning
Problem set 6

Alan Joseph Bekker

Gaussian Mixture Model

Suppose that we are given a training set x_1, .. , x_m as usual. Since we are in the unsupervised learning setting, these points do not come with any labels.We wish to model the data by specifying a joint distribution $p(x_i, z_i) = p(x_i|zi)p(z_i)$. Here z_i has a multinomial distribution .Were $p(z_i = j) = \phi_j$, $\sum_j \phi_j = 1$ and $x_i|z_i = j \sim \mathcal{N}(\mu_j, \sigma_j)$ let k denote the number of values that the z_i can take on. Thus, our model posits that each x_i was generated by randomly choosing z_i from 1, . . . , k, and then x_i was drawn from one of k Gaussians depending on z_i. This is called the mixture of Gaussians model. Also, note that the z_i are latent random variables, meaning that they are hidden/unobserved. This is what will make our estimation problem difficult. In order to solve this problem and find the z_i we apply the EM algorithm as following:

- E-step
  For each i, j, set:
  $w_i^j = p(z_i = j|x_i : \phi_j, \mu_j, \sigma_j) = \frac{p(x_i|\mu_j,\sigma_j)p(z_i=j)}{\sum_{l=1}^{k} p(x_i|\mu_l,\sigma_l)p(z_i=l)}$
- M-step

  $\phi_j = \frac{\sum_{i=1}^{m} w_j^i}{m}$
  $\mu_j = \frac{\sum_{i=1}^{m} w_j^i x^i}{\sum_{i=1}^{m} w_j^i}$

  $\sigma_j = \frac{\sum_{i=1}^{m} w_j^i (x^i-\mu_j)(x^i-\mu_j)^T}{\sum_{i=1}^{m} w_j^i}$

The assignment

Generate 2D data sampled from $K = 3$ Gaussians.

1. Determine α_1,α_2,α_3 such that $\sum_i \alpha_i = 1$
2. Determine the $\vec{\mu_i}$ and the covariance matrix $\sigma_i = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{pmatrix}$ for each Gaussian distribution in such a way the data points of each dis-

tribution are almost separated with a small overlap.

3. Generate 1000 2D data points as following
   a.For each data point i raffle a label $z_i = 1, 2, 3$ according to α_1,α_2,α_3
   b.Based on the label $z_i = l$ generate the point (x_i, y_i) from the corresponding distribution.

a. Apply K-means and GMM unsupervised algorithms in order to receive the labels of the data.
b. Compare the received labels with the real labels, which algorithm performs a better success rate?