Machine Learning Homework Sheet 11

Clustering

1 Gaussian Mixture Model

Problem 1: Consider a mixture of K Gaussians

$$p(\boldsymbol{x}) = \sum_k \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Derive the expected value $\mathbb{E}[\boldsymbol{x}]$ and the covariance $\mathrm{Cov}[\boldsymbol{x}]$.

Hint: it is helpful to remember the identity $\mathrm{Cov}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T$.

For $\mathbb{E}[\boldsymbol{x}]$ we use the law of iterated expectations

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}] &= \mathbb{E}_{p(\boldsymbol{z})}[\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z})}[\boldsymbol{x}|\boldsymbol{z}]] \\ &= \sum_{k=1}^{K} \pi_k \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z})}[\boldsymbol{x}|\boldsymbol{z}]\end{aligned}$$

plugging in the mean of multivariate Gaussian, we get

$$= \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$$

For covariance, we first compute $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$ again using the law of iterated expectations

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] &= \mathbb{E}_{p(\boldsymbol{z})}[\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z})}[\boldsymbol{x}\boldsymbol{x}^T \mid \boldsymbol{z}]] \\ &= \sum_k \pi_k \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{z})}[\boldsymbol{x}\boldsymbol{x}^T|\boldsymbol{z}]\end{aligned}$$

plugging in the covariance of multivariate Gaussian, we get

$$= \sum_{k=1}^{K} \pi_k (\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T)$$

and thus

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{x}] &= \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T \\ &= \sum_{k=1}^{K} \pi_k (\boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T) - \sum_{k=1}^{K} \sum_{j=1}^{K} \pi_k \pi_j \boldsymbol{\mu}_k \boldsymbol{\mu}_j^T\end{aligned}$$

Problem 2: Consider a mixture of K isotropic Gaussians, all with the same *known* covariances $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$.

Derive the EM algorithm for the case when $\sigma^2 \to 0$, and show that it's equivalent to Lloyd's algorithm for K-means.

We consider a GMM with identical, isotropic covariances. In that case, the responsibilities take the following form:

$$\begin{aligned}
p(z_{ik} = 1 \mid \boldsymbol{x}, \theta) &= \frac{p(\boldsymbol{x} \mid z_{ik} = 1, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(z_{ik} = 1 \mid \pi_k)}{\int p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}} && (1)\\
&= \frac{\pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} && (2)\\
&= \frac{\pi_k \exp\left(\frac{-||\boldsymbol{x}-\boldsymbol{\mu}_k||^2}{2\sigma^2}\right)}{\sum_l \pi_l \exp\left(\frac{-||\boldsymbol{x}-\boldsymbol{\mu}_l||^2}{2\sigma^2}\right)} && (3)\\
&= \frac{1}{\sum_l \frac{\pi_l}{\pi_k} \exp\left(\frac{-||\boldsymbol{x}-\boldsymbol{\mu}_l||^2 + ||\boldsymbol{x}-\boldsymbol{\mu}_k||^2}{2\sigma^2}\right)} && (4)
\end{aligned}$$

If $\boldsymbol{\mu}_k$ denotes the center that is closest to $\boldsymbol{x}$, then

$$\frac{-||\boldsymbol{x} - \boldsymbol{\mu}_l||^2 + ||\boldsymbol{x} - \boldsymbol{\mu}_k||^2}{2\sigma^2} \leq 0$$

for all l, with equality if and only if $k = l$. For $\sigma \to 0$, the denominator of Equation 4 converges to 1: If $k = l$, the argument of $\exp(\cdot)$ is exactly zero, while for $k \neq l$ we are exponentiating increasingly negative arguments.

If $\boldsymbol{\mu}_k$ denotes a center that is *not* closest to $\boldsymbol{x}$, there is at least one $l \neq k$ for which

$$0 < \frac{-||\boldsymbol{x} - \boldsymbol{\mu}_l||^2 + ||\boldsymbol{x} - \boldsymbol{\mu}_k||^2}{2\sigma^2} \longrightarrow \infty, \qquad \text{as } \sigma \to 0.$$

Consequently, the denominator of Equation 4 converges to ∞, too.

This means that the responsibilities degenerate to a hard one-hot assignment of the data point $\boldsymbol{x}$ to the component closest to $\boldsymbol{x}$. This coincides with step 1 of Lloyd's algorithm.

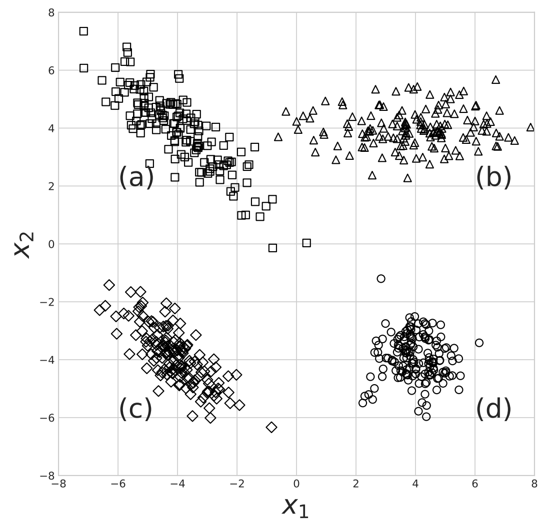
Inserting one-hot responsibilities into the general GMM M-step immediately yields step 2 in Lloyd's algorithm. Notice that we do not learn covariances, they are assumed fixed. Moreover, we don't have to worry about π_k's, because they are irrelevant as the term π_l/π_k always gets overshadowed by the $\exp(\cdot)$ next to it.

We can conclude that Lloyd's algorithm for K-Means is a special case of the more general EM algorithm for GMM.

Problem 3: The dataset displayed on the right has been generated using a Gaussian mixture model with $K = 4$ components, each with its own mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

Match the covariance matrices in the table on the left with their corresponding Gaussian components in the plot on the right. Explain each of the answers with 1 sentence.

$\mathbf{\Sigma}_k$	**Cluster**
$\begin{bmatrix} 2 & -1.7 \\ -1.7 & 2 \end{bmatrix}$	
$\begin{bmatrix} 0.9 & -0.8 \\ -0.8 & 1.2 \end{bmatrix}$	
$\begin{bmatrix} 3 & 0 \\ 0 & 0.5 \end{bmatrix}$	
$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$	



1. plot (a) - Has same shape as 2 (plot (c)), but has higher variance in all directions.
2. plot (c) - Has same shape as 1 (plot (a)), but has lower variance in all directions.
3. plot (b) - Aligned with the axes (offdiagonal entries are 0); more variance along the x_1 dimension.
4. plot (d) - Isotropic Gaussian - same variance in all directions.

Problem 4:

a) Given is the dataset displayed in the figure below. Apply the K-means algorithm to this data using $K = 2$ and using the circled points as initial centroids.

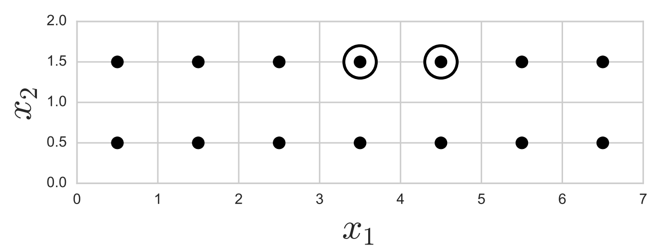


Figure 1: K-Means Dataset

What are the clusters after K-Means converges? Draw your solution in the figure above, i.e. mark the location of the centroids with $\times$'s and show the clusters by drawing two bounding boxes around the points assigned to each cluster.

How many iterations did it take for K-Means to converge in the above problem?

One.

b) Provide a different initialization, for which the algorithm will take **more** iterations to converge to the **same** solution. Make sure that your initialization does not lead to ties. Draw your initialization in the figure below.

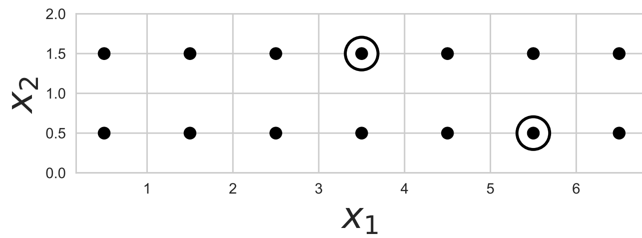


Figure 2: Provide your initialization