Exercises
Machine Learning
HS 2016

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/ml/

Series 5, Dec 2nd, 2016
(Clustering and Mixture Models)

For questions 1,2,3 : **Andreas Georgiadis**
piazza
For questions :

Please turn in solutions until Friday, Dec 9th.
("*"-exercies are a little bit more difficult, but still useful)

Problem 1 (KMeans and GMMs):

1. Assume the following data distributions. Your task in this exercise is to select the model with the **minimum model complexity** (i.e. number of free parameters) that can reasonably cluster the data.
Your options are

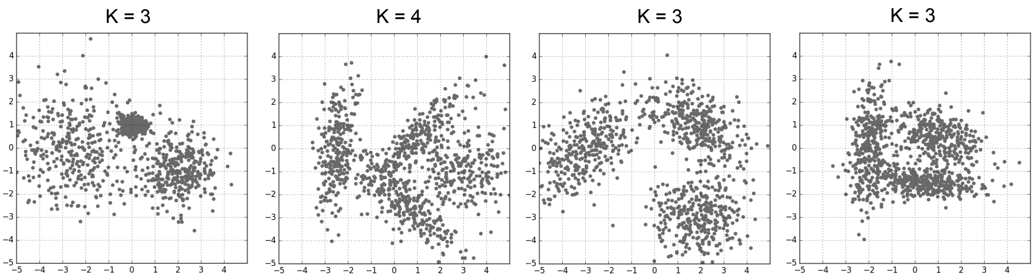
   (a) KMeans

   (b) GMMs with spherical covariance matrices $\Sigma_k = \sigma_k^2 I$

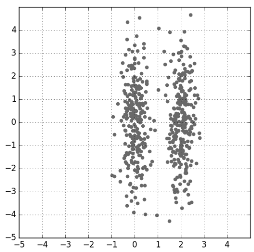
   (c) GMMs with diagonal covariance matrices $\Sigma_k = diag(\sigma_{k(1)}^2, \sigma_{k(2)}^2)$

   (d) GMMs with full covariance matrices

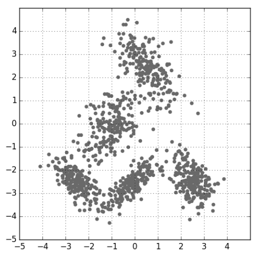
   Please clearly justify your choice. The correct number of clusters is given in each case as a hint.

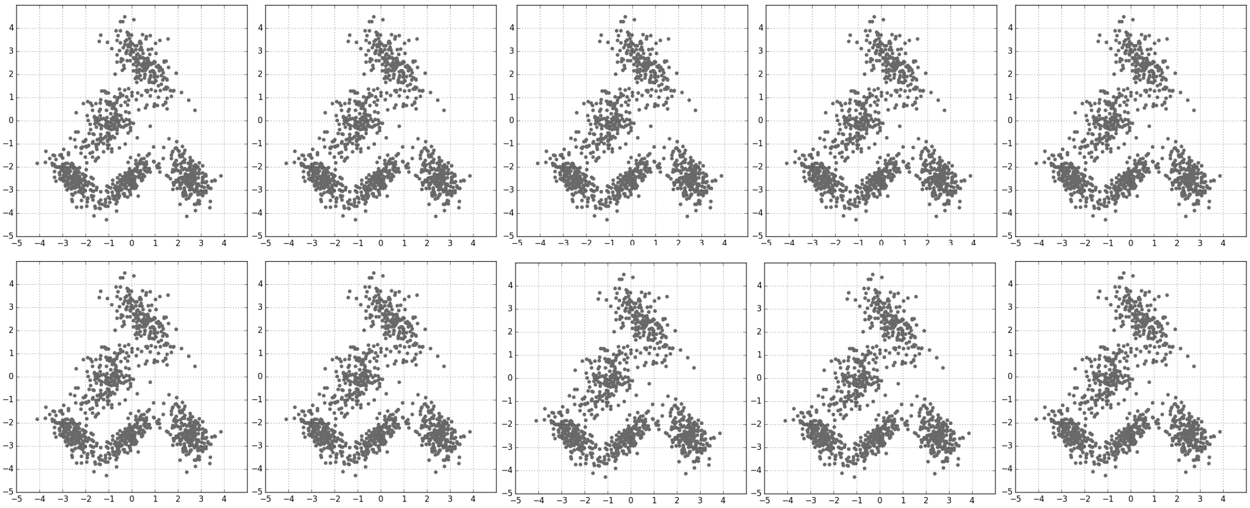


2. Assume the following data distribution. As we already mentioned in the tutorial, KMeans fails to partition the data reasonably in this case. Which kind of **data transformation** would you propose, so that the clusters would be identified correctly if we applied KMeans on the transformed data? Please clearly justify your answer. (*Hint:* More than one answers are correct)



3. Assume the following data distribution. Propose two different **cluster centroid initializations** that will lead to a different clustering outcome of the KMeans algorithm after the algorithm converges. Justify your answer by using the following plots in order to sketch the evolution of the KMeans algorithm. (*Hint:* K=5)





Problem 2 (Mixture of Poisson Distributions):

In this exercise we will apply the EM Algorithm in order to perform MLE for a mixture of Poisson distributions. The Poisson distribution models the probability of m events occuring in a time interval, when their average rate is λ. The p.m.f. of a single Poisson distribution has the form

$$\mathcal{P}(m;\lambda) = \frac{\lambda^m e^{-\lambda}}{m!},\ m \in \mathbb{N}$$

The p.m.f. of a mixture of Poisson distributions is thus given by

$$P(m;\boldsymbol{\pi},\boldsymbol{\lambda}) = \textstyle\sum_{k=1}^{K} \pi_k \mathcal{P}(m;\lambda_k)$$

Our dataset consists of i.i.d. measurements and has the form

$$\mathcal{D} = \{m_1, m_2, ..., m_N\},\ m_n \in \mathbb{N}$$

1. Derive the expression of the likelihood of the data $L(\boldsymbol{\pi},\boldsymbol{\lambda};\mathcal{D})$ as well as the log-likelihood $l(\boldsymbol{\pi},\boldsymbol{\lambda};\mathcal{D})$.
2. Introduce the latent random variables z_{kn}, which are necessary for the maximization of the log-likelihood and state their marginal distribution.
3. Calculate the posterior probabilities of the latent variables $P(z_{kn} = 1|m_n)$ when the parameters λ_k, $k = 1, ..., K$ are known.
4. Introduce the variational parameters q_{kn} in the expression of the log-likelihood and provide a lower bound by applying Jensen's inequality.
5. **E-Step:** Formulate the Lagrangian in order to maximize the lower bound with respect to the parameters q_{kn}. Prove that the optimal parameters $q^*_{kn} = \gamma_{kn}$ are equal to the posterior probabilities of the corresponding latent variables z_{kn}.
6. **M-Step:** Replace q_{kn} with γ_{kn} in the expression of the lower bound of the log-likelihood and maximize it with respect to λ_k and π_k, while keeping γ_{kn} fixed. State the expression of λ^*_k and π^*_k.

Problem 3 (Singularities in Gaussian Mixture Models):

Assume that we are given a dataset $\mathcal{D} = \{x_1, x_2, ..., x_N\}$, $x_n \in \mathbb{R}^D$ and our goal is to learn a Gaussian Mixture Model based on those observations.

1. State the expression of the log-likelihood of the mixture model given the data set, $l(\theta; \mathcal{D})$, where $\theta = (\pi_k, \mu_k, \Sigma_k)\big|_{k=1}^{K}$.

Now, consider that we constrain to a GMM with spherical covariance matrices $\Sigma_k = \sigma_k^2 I$ and that during training, a mean parameter μ_j becomes equal to one of the data points $\mu_j = x_n$.

2. State the expression of the log-likelihood for the data point x_n, $\log p(x_n; \theta)$, where $\theta = (\pi_k, \mu_k, \Sigma_k)\big|_{k=1}^{K}$.
3. Compute the likelihood of the j-th mixture component for the data point x_n, namely $\mathcal{N}(x_n; \mu_j, \Sigma_j)$.
4. What happens to the above expression as $\sigma_j \to 0$? How does this affect the expression of the log-likelihood given in question 1?
5. Can the above situation occur when the mixture model consists only of one Gaussian component?
6. Can you propose a heuristic to avoid such situations?