Exercises
Computational Intelligence Lab
SS 2017

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Prof. Dr. Thomas Hofmann
Web http://cil.inf.ethz.ch/

Series 7, April 27-28, 2017
(K-means and Mixture Models)

1 Probability Refresher and the K-means Algorithm

Problem 1 (Conditional Probability):

A couple has two children, each of them being independently a boy or a girl with 50% probability. Compute the probabilities of the following events.

1. At least one of the children is a girl.
2. Both children are girls.
3. Both children are girls given that the first born is a girl.
4. Both children are girls given that one of them is a girl.
5. Both children are girls given that one of them is a girl named Cassiopeia.
   Note: Cassiopeia is an extremely rare name with a frequency of less than 1 in 1,000,000.

Problem 2 (Bayes' Rule):

There is an uncommon disease that has infected 1% of the human population. Assume that we have a test for this disease that is positive on an infected person with probability 99% and negative on a healthy person also with probability 99%.

If my test comes out positive, what is the probability that I am infected?

Problem 3 (K-means Theory):

In this exercise, you will elaborate on some of the formal results connecting K-means theory and matrix factorization.

1. Show that the K-means algorithm always converges. In particular, consider the following cost function

$$J := \sum_{n=1}^{N} \sum_{k=1}^{K} z_{k,n} \|\mathbf{x}_n - \mathbf{u}_k\|_2^2,$$

   and show that steps 2 and 3 of the K-means algorithm from the lecture minimize this cost function for $\mathbf{z}_n$ and $\mathbf{u}_k$, respectively.

2. Show that the K-means algorithm solves a matrix factorization problem, in the sense that

$$\arg\min_{\boldsymbol{Z}} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{Z}\|_F^2 = \arg\min_{\boldsymbol{Z}} \sum_{n=1}^{N} \sum_{k=1}^{K} z_{k,n} \|\boldsymbol{x}_n - \boldsymbol{u}_k\|_2^2,$$

   when $\boldsymbol{Z} \in \mathbb{R}^{K \times N}$ is additionally restricted to be an assignment matrix (having exactly a single non-zero entry of 1 in each column). The other matrices are given as follows:

   - data matrix $\boldsymbol{X} := [\boldsymbol{x_1} \cdots \boldsymbol{x_N}] \in \mathbb{R}^{D \times N}$,
   - centroid matrix $\boldsymbol{U} := [\boldsymbol{u_1} \cdots \boldsymbol{u_K}] \in \mathbb{R}^{D \times K}$,
   - assignment matrix $\boldsymbol{Z} := [\boldsymbol{z_1} \cdots \boldsymbol{z_N}] \in \mathbb{R}^{K \times N}$.

2 Mixture Models

Problem 1 (Singularities in Gaussian Mixture Models):

In this exercise we consider the problem of singularities when maximizing the likelihood of a Gaussian mixture model. Assume we are given a data set $\boldsymbol{X}$ consisting of N i.i.d observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and our goal is to cluster these observations using a mixture of K Gaussian distributions.

1. Write down the expression for the log-likelihood of the mixture model given data $\boldsymbol{X}$ (i.e., $\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$).

Now, consider a Gaussian mixture model whose components have covariance matrices given by $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is the unit matrix and suppose that one of the components, say the j-th, has a mean parameter $\boldsymbol{\mu}_j$ that is equal to one of the data points, i.e. $\boldsymbol{\mu}_j = \boldsymbol{x}_n$ for some n.

2. Write down the expression for the log-likelihood of the mixture model given $\boldsymbol{x}_n$ (i.e., $\ln p(\boldsymbol{x}_n|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$).
3. Compute the likelihood of the j-th mixture component given $\boldsymbol{x}_n$ (i.e. $\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$).

   Hint: The multivariate Gaussian probability density function is defined as

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right).$$

4. What happens to the likelihood of the previous question as $\sigma_j \to 0$? How does this affect the log-likelihood of the mixture model given in question 1?
5. Can the above situation occur when the mixture model consists of a single Gaussian distribution, i.e. $K = 1$?
6. Can you propose a heuristic to avoid such situations?

Problem 2 (Identifiability):

In this exercise we consider the problem of *identifiability* of maximum likelihood solutions of mixture models.

1. Suppose that we have solved a mixture of K Gaussians problem and have obtained the values of the parameters. How many equivalent solutions are there?
2. This problem is known as *identifiability*. Explain why this is not a problem in the context of data clustering.