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

Problem 1 (Probability):

Let's denote the i^{th} children by a random variable X_i, for $i \in \{1, 2\}$, taking values in the set $\{girl, boy\}$. We know that X_1 and X_2 are independent, and that for all $i \in \{1, 2\}$, for all $c \in \{girl, boy\}$, $\mathbb{P}(X_i = c) = \frac{1}{2}$.

1. The probability that at least one of them is a girl is given by

   $$\mathbb{P}(\{X_1 = girl\} \cup \{X_2 = girl\}) = \mathbb{P}(\{X_1 = girl\}) + \mathbb{P}(\{X_2 = girl\}) - \mathbb{P}(\{X_1 = girl\} \cap \{X_2 = girl\}).$$

   As the events are independent, we have

   $$\mathbb{P}(\{X_1 = girl\} \cup \{X_2 = girl\}) = \mathbb{P}(\{X_1 = girl\}) + \mathbb{P}(\{X_2 = girl\}) - \mathbb{P}(\{X_1 = girl\}) \cdot \mathbb{P}(\{X_2 = girl\}),$$

   i.e.

   $$\mathbb{P}(\{X_1 = girl\} \cup \{X_2 = girl\}) = \frac{1}{2} + \frac{1}{2} - \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}.$$

2. The probability that both are girls is given by

   $$\mathbb{P}(\{X_1 = girl\} \cap \{X_2 = girl\}) = \mathbb{P}(\{X_1 = girl\}) \cdot \mathbb{P}(\{X_2 = girl\}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

3. Without loss of generality, assume that X_1 is the first born. By definition of a conditional probability, the probability that both children are girls given that the first born is a girl is given by

   $$\mathbb{P}(\{X_1 = girl\} \cap \{X_2 = girl\} | \{X_1 = girl\}) = \frac{\mathbb{P}((\{X_1 = girl\} \cap \{X_2 = girl\}) \cap \{X_1 = girl\})}{\mathbb{P}(\{X_1 = girl\})},$$

   i.e.

   $$\mathbb{P}(\{X_1 = girl\} \cap \{X_2 = girl\} | \{X_1 = girl\}) = \frac{\mathbb{P}(\{X_1 = girl\} \cap \{X_2 = girl\})}{\mathbb{P}(\{X_1 = girl\})},$$

   i.e.

   $$\mathbb{P}(\{X_1 = girl\} \cap \{X_2 = girl\} | \{X_1 = girl\}) = \frac{\mathbb{P}(\{X_1 = girl\}) \cdot \mathbb{P}(\{X_2 = girl\})}{\mathbb{P}(\{X_1 = girl\})},$$

   i.e.

   $$\mathbb{P}(\{X_1 = girl\} \cap \{X_2 = girl\} | \{X_1 = girl\}) = \mathbb{P}(\{X_2 = girl\}) = \frac{1}{2}.$$

4. The probability that both children are girls given that one of them is a girl is given by

   $$\frac{\mathbb{P}((\{X_1 = girl\} \cap \{X_2 = girl\}) \cap (\{X_1 = girl\} \cup \{X_2 = girl\}))}{\mathbb{P}(\{X_1 = girl\} \cup \{X_2 = girl\})},$$

   i.e.

   $$\frac{\mathbb{P}(\{X_1 = girl\} \cap \{X_2 = girl\})}{\mathbb{P}(\{X_1 = girl\} \cup \{X_2 = girl\})} = \frac{1/4}{3/4} = \frac{1}{3}.$$

5. Assume that the probability that children i is named Cassiopeia is equal to $p \leq \frac{1}{1{,}000{,}000}$, that only girls can be given this name, and that the random variable $name(i)$ is independent of X_j, for $i \neq j$, $i, j \in \{1, 2\}$. Then, the probability that both children are girls given that one of them is a girl named Cassiopeia is given by

$$\frac{\mathbb{P}((\{X_1 = girl\} \cap \{X_2 = girl\}) \cap (\{name(1) = Cassiopeia\} \cup \{name(2) = Cassiopeia\}))}{\mathbb{P}(\{name(1) = Cassiopeia\} \cup \{name(2) = Cassiopeia\})}.$$

Now, using that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, our probability becomes

$$\frac{\mathbb{P}((\{name(1) = Cassiopeia\} \cap \{X_2 = girl\}) \cup (\{name(2) = Cassiopeia\} \cap \{X_1 = girl\}))}{\mathbb{P}(\{name(1) = Cassiopeia\} \cup \{name(2) = Cassiopeia\})}.$$

The numerator is also equal to

$$\mathbb{P}(\{name(1) = Cassiopeia\} \cap \{X_2 = girl\}) + \mathbb{P}(\{name(2) = Cassiopeia\} \cap \{X_1 = girl\}) - \mathbb{P}(\{name(1) = Cassiopeia\} \cap \{name(2) = Cassiopeia\}),$$

which is also $p/2 + p/2 - p^2$. As the denominator is given by $p + p - p^2$, the final answer is $\frac{p/2+p/2-p^2}{p+p-p^2} = \frac{1-p}{2-p} \approx 0.49$ for $p = 1/1,000,000$.

Problem 2 (Bayes' Rule):

Let P (resp. N) denote the event being positive (resp. negative) at the test and I (resp. H) being ill (resp. healthy). We have $\mathbb{P}(P|I) = 0.99$, $\mathbb{P}(I) = 0.01$, $\mathbb{P}(N|H) = 0.99$. We want to find $\mathbb{P}(I|P)$. From Bayes' rule we have

$$\mathbb{P}(I|P) = \frac{\mathbb{P}(P|I)\mathbb{P}(I)}{\mathbb{P}(P)} = \frac{\mathbb{P}(P|I)\mathbb{P}(I)}{\mathbb{P}(P \cap I) + \mathbb{P}(P \cap H)} = \frac{\mathbb{P}(P|I)\mathbb{P}(I)}{\mathbb{P}(I)\mathbb{P}(P|I) + \mathbb{P}(H)\mathbb{P}(P|H)},$$

i.e.

$$\mathbb{P}(I|P) = \frac{\mathbb{P}(P|I)\mathbb{P}(I)}{\mathbb{P}(P)} = \frac{\mathbb{P}(P|I)\mathbb{P}(I)}{\mathbb{P}(I)\mathbb{P}(P|I) + (1 - \mathbb{P}(I))(1 - \mathbb{P}(N|H))} = \frac{0.99 \cdot 0.01}{0.01 \cdot 0.99 + (1 - 0.01) \cdot (1 - 0.99)} = \frac{1}{2}.$$

2 K-means Algorithm

Problem 3 (K-means Theory):

1. (Convergence of the K-Means Algorithm) The K-means algorithm converges since at each iteration it either reduces or keeps the same the value of the objective function J, where

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{k,n} \|\mathbf{x}_n - \mathbf{u}_k\|_2^2 \quad \left(\|\mathbf{x}_n - \mathbf{u}_k\|_2^2 = (x_{1,n} - u_{1,k})^2 + \cdots + (x_{d,n} - u_{d,k})^2\right)$$

with the constraint

$$\sum_{k=1}^{K} z_{k,n} = 1 \quad \text{and} \quad z_{k,n} \in \{0, 1\}.$$

When initializing the algorithm, at step 2 of the K-means algorithm we set

$$z_{k^*(\mathbf{x}_n),n} = 1 \quad \text{and} \quad z_{k',n} = 0,$$

where

$$k^*(\mathbf{x}_n) = \underset{k}{\operatorname{argmin}} \left\{\|\mathbf{x}_n - \mathbf{u}_1\|_2^2, \ldots, \|\mathbf{x}_n - \mathbf{u}_k\|_2^2, \ldots, \|\mathbf{x}_n - \mathbf{u}_K\|_2^2\right\}.$$

This makes the value of J minimal considering that we have to assign the value 1 to one and only one $z_{k,n}$, and 0 to all others.

At step 3, the centroid update term you are familiar with:

$$\mathbf{u}_k = \frac{\sum_{n=1}^{N} z_{k,n}\mathbf{x}_n}{\sum_{n=1}^{N} z_{k,n}} \ \forall k,\ k = 1, \ldots, K \tag{1}$$

means that

$$0 = \sum_{n=1}^{N} z_{k,n}(\mathbf{x}_n - \mathbf{u}_k) \ \forall k,\ k = 1, \ldots, K$$

Note that this equals setting the derivative of J with respect to $\mathbf{u}_k$ to zero for all k, $k = 1, \ldots, K$, as a partiuclar derivative is given by:

$$\frac{\partial J}{\partial \mathbf{u}_k} = \frac{\partial \sum_{n=1}^{N} z_{k,n}\|\mathbf{x}_n - \mathbf{u}_k\|_2^2}{\partial \mathbf{u}_k} = \sum_{n=1}^{N} z_{k,n} \begin{bmatrix} \frac{\partial (x_{1,n} - u_{1,k})^2}{\partial u_{1,k}} \\ \vdots \\ \frac{\partial (x_{d,n} - u_{d,k})^2}{\partial u_{d,k}} \end{bmatrix} = -2 \sum_{n=1}^{N} z_{k,n}(\mathbf{x}_n - \mathbf{u}_k)$$

Note that $\frac{\partial^2 J}{\partial \mathbf{u}_k^2} \geq 0$, or in other words, the gradient of J with respect to $\mathbf{u}_k$ is pointing downwards (or is flat). Thus, the value of J does not increase after the centroid update. Considering all the above, it follows that repeating steps 2 and 3 in iterations means that the value of J will converge.

2. (The K-Means Algorithm and Matrix Factorization) Notice that

$$\|X - UZ\|_F^2 = \sum_{i=1}^{D}\sum_{j=1}^{N}(x_{i,j} - \sum_{k=1}^{K} u_{i,k}z_{k,j})^2 = \sum_{i=1}^{D}\sum_{j=1}^{N}\left(\sum_{k=1}^{K} z_{k,j}(x_{i,j} - u_{i,k})\right)^2 = \sum_{k=1}^{K}\sum_{j=1}^{N} z_{k,j}^2 \sum_{i=1}^{D}(x_{i,j} - u_{i,k})^2,$$

hence

$$\|X - UZ\|_F^2 = \sum_{k=1}^{K}\sum_{j=1}^{N} z_{k,j}\|\mathbf{x}_j - \mathbf{u}_k\|_2^2.$$

3 Mixture Models

Problem 1 (Singularities in Mixture of Gaussians Models):

In this section, we study the problem of singularities in the mixture of Gaussian models. Consider the data set $\mathbf{X}$ consisting of N i.i.d observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. The goal is to cluster this data set using mixture of K Gaussians.

1. The log-likelihood of the data is given by

$$\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

2. For the data point $\mathbf{x}_n$ we have log-likelihood

$$\ln p(\mathbf{x}_n \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

3. Computing the likelihood assuming $\boldsymbol{\mu}_j = \mathbf{x}_n$ leads to

$$\begin{aligned} p(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) &= \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \\ &= \mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2\mathbf{I}) \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{\sigma_j^D} \end{aligned} \tag{2}$$

4. We see that as $\sigma_j \to 0$, (3) goes to infinity and so the *likelihood function* will also go to infinity. Thus the maximization of the log likelihood function is not a well posed problem and causes the convergence to be very slow. This can lead to a very poor clustering. Such singularities will always be present and will occur whenever one of the Gaussian components 'collapses' onto a specific data point.

5. This problem does not arise in the case of a single Gaussian distribution. If a single Gaussian collapses onto a data point it will contribute multiplicative factors to the likelihood function arising from the other data points and these factors will go to zero exponentially fast, giving an overall likelihood that goes to zero rather than infinity.

   However, once we have (at least) two components in the mixture, one of the components can have a finite variance and therefore assign finite probability to all of the data points while the other component can shrink onto one specific data point and thereby contribute an ever increasing additive value to the log likelihood.

6. We can hope to avoid the singularities by using suitable heuristics, for instance by detecting when a Gaussian component is collapsing and resetting its mean to a randomly chosen value while also resetting its covariance to some large value, and then continuing with the optimization.

Problem 2 (Identifiability):

A further issue in finding maximum likelihood solutions arises from *identifiability*. In this section we study *identifiability* in mixture models.

1. For any given maximum likelihood solution, a K-component mixture will have a total of $K!$ equivalent solutions corresponding to the $K!$ ways of assigning K sets of parameters to K components.
2. Because any of the equivalent solutions is as good as any other. Using any permutation of these parameters leads to the same clustering with permuted cluster indices.