

---

# Homework 1

---

CS420 Machine learning 2021 Spring\*  
Department of Computer Science and Engineering  
Shanghai Jiao Tong University

**Submission deadline: 20:00, April 20, 2021, Tuesday**


**Submission to:**

Please submit your homework in pdf/doc format to Canvas platform.

## 1 (10 points) k-mean algorithm

After initializing the center parameters  $\mu_1, \mu_2, \dots, \mu_K \in \mathbf{R}^n$ , the K-mean algorithm is to repeat the following two steps until convergence:

1. Assign the points to the nearest  $\mu_i$ ;
2. Update  $\mu_i$  to be the mean of the data points assigned to it.

Prove that each of the above two steps will never increase the k-mean objective function, 

$$J(\mu_1, \dots, \mu_K) = \sum_{t=1}^N \sum_{k=1}^K r_{nk} \|x_n - \mu_k\|^2, \quad (1)$$

where

$$r_{nk} = \begin{cases} 1, & \text{if } x_n \text{ is assigned to cluster } k; \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

## 2 (10 points) k-mean vs GMM

Give a variant of k-mean algorithm somewhat between the original k-mean and Expectation-Maximization (EM) for Gaussian Mixture Models (GMM). Please specify the computational details of the formulas. Pseudo-codes of the algorithm would be great.

Discuss the advantages or limitations of your algorithm.

## 3 (10 points) k-mean vs CL

Compare the k-mean algorithm with competitive learning (CL) algorithm. Could you apply the idea of Rival Penalized Competitive Learning (RPCL) to k-mean so that the number of clusters is automatically determined? If so, give the details of your algorithm and then implement it on a three-cluster dataset generated by yourself. If not, state the reasons.

---

\*tushikui@sjtu.edu.cn

#### 4 (20 points) model selection of GMM

Write a report on experimental comparisons on model selection performance between BIC, AIC and VBEM.

Specifically, you need to randomly generate datasets based on GMM, by varying some factors, e.g., sample sizes, dimensionality, number of clusters, and so on.

- **BIC, AIC:** First, **run EM algorithm** on each dataset  $X$  for  $k = 1, \dots, K$ , and calculate the log-likelihood value  $\ln[p(X|\hat{\Theta}_k)]$ , where  $\hat{\Theta}_k$  is the maximum likelihood estimate for parameters; Second, select the optimal  $k^*$  by

$$k^* = \arg \max_{k=1, \dots, K} J(k), \quad (3)$$

$$J_{AIC}(k) = \ln[p(X|\hat{\Theta}_k)] - d_m, \quad (4)$$

$$J_{BIC}(k) = \ln[p(X|\hat{\Theta}_k)] - \frac{\ln N}{2} d_m, \quad (5)$$

where  $N$  is the number of data points in the dataset,  $d_m$  is the number of free parameters in the model, and  $K$  is a positive integer specified by the user.

- Use **VBEM algorithm for GMM** to select the optimal  $k^*$  automatically or via evaluating the lower bound.

The following codes might be useful.

Matlab: <http://www.cs.ubc.ca/~murphyk/Software/VBEMGMM/index.html>

Python: <http://scikit-learn.org/stable/modules/mixture.html>