Machine Learning (Spring 2020) **(Due: 16.04.2020 23:55)**

Assignment 2

Instructor: Dr. Sergey Kosov *TA:* Yuhou Zhou

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- The homework assignments are for practice purpose. The grade from your homework will not affect your final grade of the course.
- Please submit your answer sheet, either by a scanned copy or a typeset PDF file, to Moodle before the deadline.
- No late submission is accepted.
- You can do this assignment in groups of 2. Please submit no more than one submission per group.

Problem 1: K-nearest Neighbors

(2.5+2.5+5=10 points)

(a) Given a 2 dimensional data set:

$$T = \{(2,3)^T, (5,4)^T, (9,6)^T, (4,7)^T, (8,1)^T, (7,2)^T\}$$

Construct a balanced kd-tree.

(b) Use the kd-tree constructed in problem (a) to find the nearest point of $x = (3, 4.5)^T$.

(c) Show that the k-nearest-neighbour density model defines an improper distribution whose integral over all space is divergent.

Problem 2: Gaussian Mixture Model (5+5+5+5=20 points)

(a) Consider a Gaussian mixture model in which the marginal distribution $p(\mathbf{z})$ for the latent vatiable is given by $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$, and the conditional distribution $p(\mathbf{x}|\mathbf{z})$ for the observaed variable is given by $p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$. Show that the marginal distribution $p(\mathbf{x})$, obtained by summing $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ over all possible values of $\mathbf{z}$, is a Gaussian mixture of the form $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Note: $\mathbf{z}$ uses 1-of-K representation.

(b) Verify that maximization of the complete-data log likelihood

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

for a Gaussian mixture model leads to the result that the means and covariances of each component are fitted independently to the corresponding group of data points, and the mixing coefficients are given by the fractions of points in each group.

(c) Show that if we maximize

$$\mathbb{E}_{\mathbf{z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

with respect to $\boldsymbol{\mu}_k$ while keeping the responsibilities $\gamma(z_n k)$ fixed, we obtain the closed form solution given by

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

(d) Consider a density model given by a mixture distribution

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k)$$

and suppose that we partition the vector $\mathbf{x}$ into two parts so that $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$. Show that the conditional density $p(\mathbf{x}_b|\mathbf{x}_a)$ is itself a mixture distribution and find expressions for the mixing coefficients and for the component densities.