Assignment 1

Eldor Fozilov
20192032
UNIST
eldorfozilov@unist.ac.kr

March 2022

1 K-means algorithm

The k-means algorithm aims to divide a given set of observations into a user-defined number of k clusters. All observations x are allocated to their nearest center-point during each update step (see equation (1)).

$$S_i^{(t)} = \{x_p : \left\|x_p - \mu_i^{(t)}\right\|^2 \leq \left\|x_p - \mu_j^{(t)}\right\|^2 \; \forall j, 1 \leq j \leq k\} \tag{1}$$

2 Maximum Likelihood Estimate [10pt]

The likelihood function is nothing but a parameterized density $p(D\,|\,\boldsymbol{\theta}\,)$ that is used to model a set of data $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ which are assumed to be drawn independently from $p(D\,|\,\boldsymbol{\theta}\,)$:

$$p(D\,|\,\boldsymbol{\theta}\,) = p(\boldsymbol{x}_1\,|\,\boldsymbol{\theta}\,) \;\cdot\; p(\boldsymbol{x}_2\,|\,\boldsymbol{\theta}\,) \;\cdot\; ... \; p(\boldsymbol{x}_n\,|\,\boldsymbol{\theta}\,) = \prod_{k=1}^{n} p(\boldsymbol{x}_k\,|\,\boldsymbol{\theta}\,) \tag{2}$$

Maximum likelihood seeks to find the optimum values for the parameters by maximizing a likelihood function form the training data. The log-likelihood is given by

$$l(\theta) = \sum_{k=1}^{n} ln \; p(x_k|\theta) \tag{3}$$

3 Gaussian distribution [10pt]

A Gaussian distribution is a type of continuous probability distribution for a real-valued random variable. Univariate Gaussian distribution that is of the form

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{4}$$

Multivariate Gaussian distribution that is of the form

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{5}$$

4 Estimation method derivation [10pt]

Please show the estimation results of maximum likelihood for $p(x) \sim N(\mu|\sigma^2)$, where both μ and σ^2 are unknown.