

Assignment for Lecture 11

CLUSTERING, ROBUSTNESS AND SPARSITY

Lecture Date: 5/13/2026

“C” denotes for “computational” problems, language suggestion: Python/Julia

please include codes and results with analyses for computational problems

please write in pdf format and submit to bjcai@fudan.edu.cn before the lecture of 5/20/2026

1. Show that the GMM approaches to the k -means clustering in the limit $\epsilon \rightarrow 0$, where ϵ is the width of the Gaussian.
2. Denote $\mathbf{t} = m^{-1} \sum_{i=1}^m \mathbf{x}^{(i)}$ be the cluster center of a series of unit vectors $\mathbf{x}^{(i)}$'s and define the similarity between two points $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ as their dot product. Prove that the average cluster similarity $m^{-2} \sum_{i,j} \mathbf{x}^{(i)} \mathbf{x}^{(j)\top}$ is the same whether it is computed by averaging all pairs or by computing the average similarity of each point with the centroid of the cluster.
3. [C] Tukey's loss function is defined as

$$\chi^{\text{Tuk}}(e) = \begin{cases} [1 - (1 - e^2/\eta^2)^3] \eta^2/6, & |e| \leq \eta, \\ \eta^2/6, & |e| > \eta, \end{cases} \quad (1-1)$$

where η is the threshold. Apply Tukey's loss function to do the learning process for the data $y^{(i)} = x^{(i)} + \text{rand} \times \delta$ under the learning model $f_{\mathbf{w}}(x) = w_0 + w_1 x$.