CSCI8000 - HW1

Instructor: Dr. Ninghao Liu (ninghao.liu@uga.edu)

August 27, 2021

– Upload two files to eLC:
1) a scanned handwritten solution or typed pdf file named "*YourID_HW1.pdf*" containing your answer to each question;
2) a zip file named "*YourID_HW1PQ.zip*" containing your programs for Question 5.
(Note: The output of program for each subquestion in Question 5 should be included in the pdf file "*YourUIN_HW1.pdf*".)
– Due Date: September 10, 2021

1 Statistical Methods (5pt + 5pt = 10pt)

(1) In Kernel density estimation, given samples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ and a test instance $\boldsymbol{x}$,

$$p(\boldsymbol{x}) = \frac{1}{Nh} \sum_{i=1}^{N} K(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}) \tag{1}$$

where $K()$ is a valid kernel function, and h is the bandwidth parameter. Prove that $\int_{-\infty}^{+\infty} p(\boldsymbol{x}) d\boldsymbol{x} = 1$.
(2) When using Kernel density estimation for outlier detection, explain the steps of how to use validation to choose the best bandwidth parameter value h.

2 LOF (10pt)

Given a dataset of 5 data points $\{[1,1],[3,3],[4,3],[3,4],[5,3]\}$, let $\boldsymbol{p} = [1,1]$. If we set $k = 2$,

- Compute the k-distance of $\boldsymbol{p}$.
- What is the k-distance neighborhood of $\boldsymbol{p}$?
- Let $\boldsymbol{o} = [3,4]$, what is the reachability distance $reach\_dist_k(\boldsymbol{p}, \boldsymbol{o})$.
- Let $MinPts = 2$, what is the local reachability density of $\boldsymbol{p}$?
- Compute the LOF score of $\boldsymbol{p}$.

3 Clustering Based Methods (10pt + 10pt = 20pt)

(1) Given a set of observations $(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$, k-means clustering aims to partition the N observations into $k(\leq n)$ sets $S = \{S_1, S_2, ..., S_k\}$. Formally, the objective is to find:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in S_i} \|\boldsymbol{x} - \boldsymbol{\mu_i}\|^2, \tag{2}$$

where $\boldsymbol{\mu_i}$ is the mean vector of S_i. Prove that this is equivalent to the objective below:

$$\arg\min_{S} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\boldsymbol{x}, \boldsymbol{y} \in S_i} \|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{3}$$

(2) Run the k-means algorithm on the dataset $[[1,1],[5,5],[5,6],[1,2],[6,6],[2,1]]$ with $k = 2$ and initial mean vectors as $\boldsymbol{\mu}_1 = [5,4]$ and $\boldsymbol{\mu}_2 = [5,7]$. Please show the details of assignment and means update in each iteration.

4 Fundamental Optimization (10pt)

Apply the KKT conditions to solve the problem below:

$$\begin{aligned} &\max_{x,y} \ 4x^2 + 10y^2 \\ &s.t., \ x^2 + y^2 \leq 2 \end{aligned} \tag{4}$$

5 Programming Questions (50pt)

5.1 Outlier Detection with Gaussian Distribution (10pt)

In this problem, we will detect outliers, using statistical methods, from a dataset with 600 instances, where each instance has 2 features. Assume that the data is generated by a Gaussian distribution. Return the top 3 outliers (report their coordinates), and the mean vector and the covariance matrix. The dataset could be found in the file "data_1.npy", which could be opened as below.

```
import numpy as np

with open('data_1.npy', 'rb') as f:
    X = np.load(f)
```

5.2 LOF (20pt)

In this problem, we will detect outliers using LOF still on the "data_1.npy" dataset. The base codes could be found in "lof.py". Complete the LOF algorithm in "lof.py". Return the top 3 outliers (report their coordinates).

5.3 Isolation Trees (20pt)

In this problem, we will detect outliers using Isolation Forest on the "data_2.npy" dataset. The base codes could be found in "iForest.py". Complete the Isolation Forest algorithm in "iForest.py". Return the top 4 outliers (report their coordinates).