Assignment 4

October 2021

Instructions

- This assignment should be completed individually.
- Do not look at solutions to this assignment or related ones on the Internet.
- The files related to the assignment are present in `lab4-rollno.zip` folder. Extract it and upload it on moodle in the same .zip format after completion and after replacing the string "rollno" with your actual roll number. For example, if you roll number is 00405036, then single zip folder that you will upload will be named "lab4-00405036.zip". Also collate all the CS337 based theory solutions into ONE pdf file named `answers.pdf`. Include `answers.pdf` inside the zip folder mentioned above and submit the zip folder.
- Answers to all subjective questions need to be placed in single pdf `answers.pdf` including all plots and figures and uploaded.
- Only add/modify code between `TODO` and `END TODO` unless specified otherwise. You must not import any additional libraries.
- This Assignment carries a total of **8** marks for CS337 Theory and **14.5** marks for CS335 Lab

1 Clustering

1.1 CS 335 KMeans Implementation

(i) In `assignment_4.ipynb`, complete the functions `fit`, `predict` of `Kmeans` class. (2 marks)

(ii) Report the clusters formed by the KMeans algorithm on the datasets given the dataset as a numpy matrix `assignment_4.npy` with 3 different seed values. Comment on the quality of the cluster returned. (2 marks)

(iii) What would be a good initialisation for the K-means algorithm? Briefly justify your choice. (2 marks)

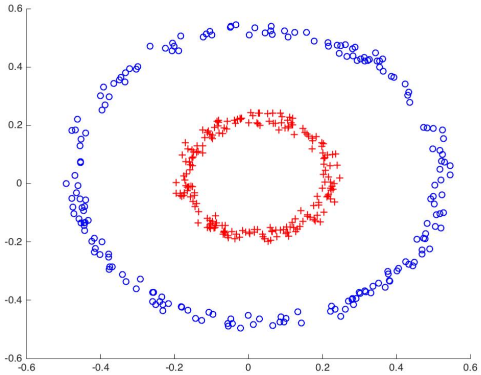

Figure 1: Concentric Circles Dataset

2 Kernel design and Kernelized clustering

2.1 CS 337: Proving kernel validity

Prove that the function $K_\sigma : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined as $K_\sigma(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(\frac{-\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right)$ is a valid Kernel. You may use the properties proved in class. [Hint - Taylor series expansion] (2 marks)

2.2 CS 337: Simple Kernel Design

Figure 1 shows a zero-centered dataset where blue points have label 0 and red points have label 1. Each blue point is at a distance $r_1 \pm \epsilon_1; \epsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$ from the origin and each red point is at a distance $r_2 \pm \epsilon_2; \epsilon_2 \sim \mathcal{N}(0, \sigma_2^2)$ from the origin.

(i) Is there a condition on r_1, r_2 such that the vanilla KMeans (vanilla means we need to run the algorithm as is on the given data without transformations of any kind) algorithm gives us the correct clusters? Explain with sound arguments. (2 marks)

(ii) For the configurations of r_1, r_2 that are not clusterable, can you suggest a kernel that will help KMeans identify the correct clusters? Specify both the transformation $\phi(x)$ and the kernel function $k(x, x\prime)$. Further show that the kernel function you propose is a valid kernel. (3 marks)

You will implement KMeans using the kernel function you identify here in your programming assignment by completing the functions `fit`, `predict` of `Kmeans_Kernel` class. First you have to complete `make_zero_centered` helper function which makes any concentric circles dataset zero centered.