CMU 10-715: Homework 9

Linkage Based Clustering

DUE: Nov. 24, 2020, 11:59 PM.

Instructions:

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books, again after you have thought about the problems on your own. Please don't search for answers on the web, previous years' homeworks, etc. (please ask the TAs if you are not sure if you can use a particular reference). There are two requirements: first, cite your collaborators fully and completely (e.g., "Alice explained to me what is asked in Question 4.3"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only.
- **Submitting your work:** Assignments should be submitted as PDFs using Gradescope unless explicitly stated otherwise. Each derivation/proof should be completed on a separate page. Submissions can be handwritten, but should be labeled and clearly legible. Else, submission can be written in LaTeX.
- **Late days:** For each homework you get three late days to be used only when anything urgent comes up. No points will be deducted for using these late days. We will consider an honor system where we will rely on you to use the late days appropriately.

1 Linkage Based Clustering

In this homework you will implement a Linkage Based Clustering algorithm, you will try variants of the clustering algorithm by changing the distances of the original space and the distances defined for the clusters. Your main objective will be to apply the unsupervised algorithm to see if the "unsupervised" labels produced by your implementation match the "true" labels of the dataset provided for the homework.

Since this is an unsupervised algorithm the evaluation will be solely on the report of your findings and not on the performance of the algorithm. Please do not use the "true" labels that are provided to you until the end, once you feel confident that your algorithm has achieved reasonable results.

We first describe single-linkage algorithm. Let the training matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. For easier exposition we will denote as $D : \mathcal{C} \times \mathcal{C} \mapsto \mathbb{R}$ the distance function between a pair of clusters, and $d : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ the distance function between observations of the original space.

Algorithm 1: Linkage Based Algorithm

Input: Training matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, cluster distance D and number of clusters k. *

Begin with disjoint clustering (each point considered a cluster).

While number of clusters larger than k:

1. Find the pair of clusters $\mathcal{C}_{i^*}, \mathcal{C}_{j^*}$ with the minimum distance:
$$\mathcal{C}_{i^*}, \mathcal{C}_{j^*} = \underset{\mathcal{C}_i, \mathcal{C}_j \in \mathcal{C}}{\operatorname{argmin}} \left(D(\mathcal{C}_i, \mathcal{C}_j) \right)$$
2. Merge the clusters into $\mathcal{C}_{new} = \mathcal{C}_{i^*} \cup \mathcal{C}_{j^*}$.

Output: Cluster assignation

*Note: the cluster distance D incorporates the selection of d.

1.1 Instructions

- Download the data from `https://github.com/ShenghaoWu/10715/tree/master/hw9` and read the train matrix available at data/train.csv. (Use the data/labels.csv for the true labels at the end). Your goal is to perform clustering on this dataset.
- Code up the single-linkage clustering algorithm.
- Your goal is to design and choose some distance functions (more details below) so that based on the *training set* you think you have a good clustering algorithm for this data. You can feel free to do whatever you want with the training data, e.g., plot it in any way you want or eyeball the individual datapoints etc. In this process, get intuition behind the data and then either choose from below or design your own distance functions.
- Once you choose/design your distance function, you set that in stone and then look at the test data. You will be asked to report the performance of the clustering algorithm with your chosen distance function on the test data. Please keep in mind that our overall bar for a good grade in this homework is very low, so please abide by a honor code of not looking at the test data, and don't worry too much if you don't get a great performance on the test data. The goal is to get your hands dirty and have a good learning experience.
- You will need to use some distance function, possibly create your own. Here are some examples of popular distance functions. You can feel free to try out these distances (this is completely optional) before you start thinking about what distance you eventually want to choose/design and use:
  1. Euclidean Distance $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\intercal (\mathbf{x}_i - \mathbf{x}_j)}$
  2. Cosine Distance $d(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\mathbf{x}_i^\intercal x_j}{||\mathbf{x}_i|| \mathbf{x}_j||}$
  3. City Block Distance $d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{h=1}^{d} |x_{i,h} - x_{j,h}|$
- Select among the following variants of the Linkage based algorithm by changing the clusters distance $D : \mathcal{C} \times \mathcal{C} \mapsto \mathbb{R}$ accordingly. You can feel free to pick any one and stick with it or try all three and pick the one which you feel works best – your choice.
  1. Single linkage $D(\mathcal{C}_h, \mathcal{C}_k) = \min_{\mathbf{x}_i \in \mathcal{C}_h, \mathbf{x}_j \in \mathcal{C}_k} (d(\mathbf{x}_i, \mathbf{x}_j))$
  2. Max linkage $D(\mathcal{C}_h, \mathcal{C}_k) = \max_{\mathbf{x}_i \in \mathcal{C}_h, \mathbf{x}_j \in \mathcal{C}_k} (d(\mathbf{x}_i, \mathbf{x}_j))$
  3. Average linkage $D(\mathcal{C}_h, \mathcal{C}_k) = \sum_{\mathbf{x}_i \in \mathcal{C}_h} \sum_{\mathbf{x}_j \in \mathcal{C}_k} \frac{d(\mathbf{x}_i, \mathbf{x}_j)}{|\mathcal{C}_h| * |\mathcal{C}_k|}$
- We recommend you to try dimensionality reduction techniques as a preprocessing step. Explore the **train** data to make this decision and feel

free to play with multiple functions before deciding the final one! Again, it is important to note that for this part you are not allowed to look at the labels set.

1.2 Results

a (50 points) Report the distances that you used and the intuition behind your selection. This includes your selection for the distance between points and the distance for the clusters. Describe in detail the process that you went through for the selection of the distances.

b (50 points) Create a scatter plot with the cluster labels that you obtained and compare it with a scatter plot with the true labels. For this scatter plot use Principal Components Analysis and plot the first two components.