CS4786/5786: Machine Learning for Data Science, Fall 2016
09/22/2016: Assignment 3: Single Link and K-means Clustering

Instructions Due at 11:59pm September 29th on CMS. Submit what you have at least once by an hour before that deadline, even if you haven't quite added all the finishing touches — CMS allows resubmissions up to, but not after, the deadline. If there is an emergency such that you need an extension, contact the professor. You have a slip day of at most one day for the assignment. The assignment is to be done individually. You will submit both a writeup and some datafiles you create. The writeup can be handwritten or typeset, but please make sure it is easily readable either way. Keep an eye on the course webpage for any announcements or updates.

Academic integrity policy We distinguish between "merely" violating the rules for a given assignment and violating *academic integrity*. To violate the latter is to commit *fraud* by claiming credit for someone else's work. For this assignment, an example of the former would be getting an answer from person X but stating in your homework that X was the source of that particular answer. You would cross the line into fraud if you did not mention X. The worst-case outcome for the former is a grade penalty; the worst-case scenario in the latter is academic-integrity hearing procedures.

The way to avoid violating academic integrity is to always document any portions of work you submit that are due to or influenced by other sources, even if those sources weren't permitted by the rules.[1]

[1] We make an exception for sources that can be taken for granted in the instructional setting, namely, the course materials. To minimize documentation effort, we also do not expect you to credit the course staff for ideas you get from them, although it's nice to do so anyway.

*Q1 (***Clustering Sensitivity).** In class, we covered k-means and single link clustering methods. The goal of this assignment is for you to explore the sensitivity of the clustering methods we've introduced by showing that small perturbations of the initial data can lead to a quite different clustering, even for binary clusterings.

You may use code packages provided by other people or sources — be sure to credit these sources appropriately. But, *if you use external code, it is your responsibility to make sure that the resulting clusterings are the same as would be produced if you were to reimplement precisely what was presented in class.* For example, if you use k-means clustering code that makes multiple runs and then averages over the runs in some way[2], then your clustering result may differ from what our testing harness comes up with. Thus, carefully read the documentation of any external code. Specifically for k-means, if you are using external code, make sure to specify explicitly initial centroids (usually this can be added as extra parameter).

In grading, we care about your explanations at least as much as the datasets you provide.

In this assignment, K, the number of clusters per clustering, is fixed at 2, and n, the number of data points each initial dataset should contain, is fixed at 30. When you are asked to provide a vector c of cluster assignments, in such vectors, the t^{th} entry c_t is 1 if the t^{th} datapoint is in the first cluster, 0 otherwise.[3]

Q 1.1 **K-means:**

- Create an initial data matrix $X^{\text{kmeans},I}$ with 30 points each in $\mathbb{R}^2$. Also create the vector $c^{\text{kmeans},I} \in \mathbb{R}^{30}$ of cluster assignments you get by running the K-means algorithm on this data along with the initial two cluster centers $\mu_1, \mu_2 \in \mathbb{R}^2$ you chose to use. $c^{\text{kmeans},I}$ **should have an equal number of** 1**'s and** 0**'s, that is, clusters of equal size.**
- Add anywhere between 1 to 3 points to $X^{\text{kmeans},I}$ to create a new data matrix $X^{\text{kmeans},II}$. **These** 1 **to** 3 **points must be within the smallest rectangle bounding the points in** $X^{\text{kmeans},I}$**, and must be the last vectors in your matrix.** Run the K-means algorithm on this modified dataset with the **same initial cluster centers** μ_1, μ_2 you used for $X^{\text{kmeans},I}$ and produce the new cluster assignment vector $c^{\text{kmeans},II}$.
- **Goal:** $c^{\text{kmeans},II}$ **and** $c^{\text{kmeans},I}$ **must vary by over 30%. That is**[4]**,**
  $\min_{C=c^{\text{kmeans},II}, C=\mathbf{1}-c^{\text{kmeans},II}} \frac{1}{30}\sum_{t=1}^{30} \mathbb{1}_{\{c_t^{\text{kmeans},I} \neq C_t\}} \geq 0.3$

Q 1.2 **Single Link:**

- Create an initial data matrix $X^{\text{s-link},I}$ with 30 points each in $\mathbb{R}^2$. Also create the vector $c^{\text{s-link},I}$ of cluster assignments you get by running the single link clustering algorithm on it. $c^{\text{s-link},I}$ **should have an equal number of** 1**'s and** 0**'s.**

[2] Hint: we didn't just make this up.

[3] It's up to you which cluster is the "first" one, so in this sense the cluster labels are arbitrary; we just need to know which points are in different clusters and which points are in the same cluster.

[4] $\mathbf{1}$ is the vector with all 30 coordinates being 1. We pick C this way because labeling clusters as $1-0$ or $0-1$ leads to the same groupings but potentially swapped labels, so just looking at label differences isn't the right way to measure the degree of perturbation. So this measure checks both one labeling and then the "flip" of that labeling.

- Add anywhere between 1 to 3 points to $X^{\text{s}-\text{link},I}$ to create the data matrix $X^{\text{s}-\text{link},II}$. **These 1 to 3 points must be within the smallest rectangle bounding the points in $X^{\text{s}-\text{link},I}$, and must be the last vectors in your matrix.** Run the single link clustering algorithm on this modified dataset and produce the new cluster assignment $c^{\text{s}-\text{link},II}$.
- **Goal: $c^{\text{s}-\text{link},II}$ and $c^{\text{s}-\text{link},I}$ must vary by over 30%. That is,** $\min_{C=c^{\text{s}-\text{link},II},C=\mathbf{1}-c^{\text{s}-\text{link},II}} \frac{1}{30}\sum_{t=1}^{30} \mathbb{1}_{\{c_t^{\text{s}-\text{link},I} \neq C_t\}} \geq 0.3$

Deliverables: Submit a **writeup** explaining the way you generated the data points and the corresponding modifications, and why you expected the new datasets to result in significantly different clusterings. In your write-up, for every cluster in the final (output) clustering produced, **include scatter plots** of the points where the points are color-coded according to their corresponding cluster assignments. For K-means, also include your **initial cluster centroids (as larger points or otherwise clearly visible and distinguished from the data points)** in the scatter plots.

For each method, also submit the initial data points; the modified dataset matrix produced by adding the extra 1 to 3 points (or edges); and the cluster assignments you obtained by running the algorithms over the initial and modified datasets. For the K-means algorithm provide the initial cluster means μ_1, μ_2 you started with.

Specifically, submit your datasets as csv files obeying the following requirements. `XkmeansI.csv` and `XslinkI.csv` must each consist of exactly 30 lines, each consisting of 2 comma-separated values. `XkmeansII.csv` and `XslinkII.csv` must each be between 31 and 33 lines, where each line contains 2 comma-separated values.

Finally, `ckmeanI.csv`, `ckmeanII.csv`, `cslinkI.csv`, and `cslinkII.csv` are each 30 lines, where each line contains one value that is either 0 or 1, indicating the cluster assignment of the corresponding original point. Also submit cluster centers $\mu_1, \mu_2 \in \mathbb{R}^2$ for Q1.1 in file `means.csv` containing 2 lines, representing μ_1 and μ_2, respectively, each of which consists of 2 comma-separated values.

Note: in this assignment, points will be deducted for submissions of dataset that do not conform precisely to our instructions. (Last time, we altered our grading code to handle transposes and the like).