**CS7641 Unsupervised Learning HW3**

Aleksandr Plokhikh ([aplokhikh3@gatech.edu](mailto:aplokhikh3@gatech.edu))

**Introduction, dataset selection, data preprocessing**

The goal of the assignment is study unsupervised learning algorithms such as clustering, and feature transformation + selection, and study how the last can help with supervised learning (NN), and how substitution of the labels with the ones from clustering affects on supervised learning and possible fixes preexisted mislabeling.

For the task I take the same fruits, and phones dataset which are interested for a number of reasons which described in A1, but here for A3 we have a new dimention of interest. Fruits have pretty determined classes of different fruits with genes of the fruits, but in case of phones, the cell phones just divided on 4 price ranges, and we do not know how arbitrary is the devision on clasees, we can easily do more classes, and do not break any logic, but it will be nice to se what is the reasonable number of price ranges(clusters there).

I did data preprocessing fro both datasets (linearly scaled all features from 0 to 1 where 0 -is the minima featurel value in the set, and 1 – is the maximum). This is extremely important step for A3. It is very important for determining distancies for KMeans, reasonable expectation calculations for EM, and very important for some feature transformation techniques especially crucial for PCA! For example for phones dataset variance of the phones in battery capacity in mAh is 2-3 orders magnitude higher than difference in CPU clock speed in GHz, and there is obvious difference in importance for customer. For example everage consumer may not see difference between phones with 3502mAh and 3500mAh difference but corresponding difference between 3.5Ghz, and 1.5GHZ of CPU clok speed will be visible as day and night especially if we take into account that phone with 3.5Ghz CPU most likely will have more cores, and definitely different architecture with wider pipelines which are not usually listed in phones spec list but have a very substantial multiplier on performance even further than simple ratio of 3.5Ghz/1.5GHz.

If we will not implement scaling then PCA on phones dataset will decide that principal components are battery capacity in mAh, mass of phones in grams, and length of it in mm or some linerar combination of them since there are obvioce positive covariance of this parameters. They will dominate the process while they are not the most crucial features. Clustering will mbe also spoiled and will divide phones on clusters based on battery capacity, weight, and length ignoring other aspect with euqlidian or Manhattan distancing in Kmeans or in EM. The same situation will be in fruits datset but with other parameters like mass, and length of the fruit. Based on my knowledge I can parse reasonable scaling of the parameters but it ruins the idea unsupervised learning, I can not pass damain knowledge to the algorithm, and just do the best most simple knowledge independent scaling of each parameter from 0 to 1.

Part 1, Clustering

On fruits dataset most simple unsupervised standard square error (SSE) performance measure method show that with increasing of number of clusters for KMeans method of clustering gradually decreases, which does make perfect sense. The plot does not show any noiticible inflection point. SSE is not good method for the database. Unsuppervised Silhoutte method on contrary shows very prominent inflection on number of clusters equal to 11 for KMeans, and 10 for EM. Any further increase in clusters number does not make sence especially for EM algorithm since the metric does not go lower anymore. The supervised homogeneity score does show inflection at number of clusters equal to 7 (the real number of clusters) for both methods, the homogeneity score is also high overall which says that both clustering method produces clusters cvery close to the right labes. The larger number of clusters offered by Sihoutte method may say that for larger fruit types (linear size and wheigh) very is a some variation itraclass variation in size which causes etra clustering. Small variation of big values in one class for KMeans, and EM are as significant as many variation of smal parameter in different classes. It wmast be the issue of datascaling but I did not have a right to offer better scaling method without parsing additional domain of knowledge to the algorithm.