# 2IMW30 - Foundations of data mining

Quartile 3, 2015-2016

Assignment 1b (10 Points + 4 Bonus points)

Deadline: Thursday 25 February (noon)

In this assignment you will experiment with different clustering algorithms and prove correctness of clustering algorithms. Let $P = \{p_1, \ldots, p_n\}$ be a set of $n$ points from $\mathbb{R}^d$. Recall from the lecture that the $k$-means problem belongs to the class of facility location problems. For these problems, a solution always consists of a set of $k$ centers (points) which we denote by $C = \{c_1, c_2, \ldots, c_k\}$. Each input point $p_i$ is assigned to its nearest neighbor in $C$, namely the point that realizes $\text{argmin}_{c_j \in C} \|p_i - c_j\|$. The centers partition the input points into $k$ clusters, each cluster being the set of points of $P$ which were assigned to its center.

**Exercise 1 (6 Points)**

Implement Lloyd's algorithm ($k$-means) with different methods for initialization, choosing the first $k$ centers as follows

(a) the first $k$ points of the data set,

(b) $k$ points of $P$ picked uniformly at random,

(c) implement the $k$-means++ algorithm (see lecture) and use this as initialization,

(d) implement Gonzales' algorithm (see lecture) and use the $k$ centers that were computed by your implementation (to avoid too much variation in the results, use the first point of the dataset as the first center $c_1$).

Perform experiments for different values of $k = \{3, 4, 5\}$ with the dataset C2.txt provided through canvas. Note that (b) and (c) need to be run several times in order to be evaluated, since they use randomization. Report for each fixed $k$ the development of the $k$-means cost across the iterations of the algorithm (for the initializations that use randomization plot a sample of 5 runs). Compare the final costs resulting from the different initializations, for (b) and (c) report the average and standard deviation of the final costs of the different runs. Supplement your report with images of the clusterings (e.g., by showing each cluster in a different color).

**Exercise 2 (3 Points)**

Prove that the update step of Lloyd's algorithm computes an optimal solution to the 1-mean problem for each cluster of the current partitioning. Recall that the update step computes for each cluster $U \subseteq P$ the barycenter point

$$b := \frac{1}{n} \sum_{p_i \in U} p_i.$$

You should prove that this barycenter minimizes

$$\phi(U, b) = \sum_{p_i \in U} \|p_i - b\|^2.$$

(Hint: Assume first that $d = 1$.)

**Exercise 3 (1 Point)**

The cosine distance between two vectors $p, q \in \mathbb{R}^d$ is defined using the cosine similarity as follows

$$\mathrm{d}_{\cos}(p,q) := 1 - \mathrm{sim}_{\cos}(p,q) = 1 - \frac{\cos(\alpha)}{\|p\|\|q\|}$$

Does this distance function satisfy the triangle inequality? Prove your answer.

**Exercise 4 (1+3 (Bonus) Points)**

In this exercise you may experiment with another variant of the clustering problem that we did not discuss in the lecture. This exercise is open-ended, any approach is possible.

The $k$-median problem is a facility location problem, where the cost function

$$\phi(P,C) = \sum_{p_i \in P} \|p_i - (\operatorname*{argmin}_{c_j \in C} \|p_i - c_j\|)\|$$

is to be minimized.

(a) Note the similarity to the $k$-means problem. Which of the two do you expect to be more robust against noise? Explain your answer.

(b) Design an algorithm for this optimization problem (to find a solution which is not necessarly optimal, but as good as possible). Implement your algorithm and compute a set of 4 centers $C = \{c_1, c_2, c_3, c_4\}$ for the dataset C3.txt provided through canvas. Report the centers as well as $\phi(P,C)$. Briefly describe your algorithmic approach. Your score will be based on how small a $\phi(P,C)$ you can find.