Data Mining

Homework 3

Due: 9/12/2018, 23:59

Instructions

You must hand in the homework electronically and before the due date and time.

This homework has to be done by each **person individually**.

Handing in: You must hand in the homework by the due date and time by an email to Leonardo (`martini.1722989@studenti.uniroma1.it`) that will contain as attachment **(not links to some file-uploading server!)** a .zip file with your answers. The filename of the attachment should be `DM_Homework_1__StudentID_StudentName_StudentLastname.zip`;
for example:
`DM_Homework_1__1235711_Robert_Anthony_De_Niro.zip`.
The email subject should be
`[Data Mining] Homework_1 StudentID StudentName StudentLastname`;
For example:
`[Data Mining] Homework_1 1235711 Robert Anthony De Niro`.
After you submit, you will receive an acknowledgement email that your project has been received and at what date and time. If you have not received an acknowledgement email within 2 days after you submit then contact Leonardo.

The solutions for the theoretical exercises must contain your answers either typed up or hand written clearly and scanned.

For information about collaboration, and about being late check the web page.

Problem 1. In this question you have to cluster the *Kijiji* announcements you downloaded in the previous homework. We are asking you to cluster them using different techniques. This problem is divided in several steps:

1. First of all, you have to pre process each document. You can do whatever preprocessing you think is essential such as stop word removal, normalization and stemming. Then, each announcement has to be represented as a vector of $Tf \cdot Idf$.

2. The Next step is to cluster these documents using two different approaches:

   (a) Apply directly a clustering algorithm to the original datasets

   (b) Apply SVD and then cluster the data using the truncated matrix as input.

   Notice that the metrics you have to use to cluster the data is the **cosine similarity** defined as:

$$d(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

3. Compare the outputs of the two approaches with respect to *Davies-Bouldin index*.

4. Here we want to explore the clusters that we have created. Look at the clusters and try to explain the announcements in the various clusters. It may help to try to find the words that characterize each cluster. It is up to you to design an algorithm for selecting the most

descriptive words for each cluster. Not that for this question there does not exist a single correct approach, so try different approaches and justify them. You may also7 want to experiment with different values of k.

Problem 2. We will now study some questions of k-means on 1 dimension.

1. Recall that in the k-means problem we want to minimize the total squared ℓ_2 distance between each point and the center to which it is assigned to:

$$\sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2,$$

where C_i is the set of points that belong to the ith cluster, $\boldsymbol{\mu}_i$ the mean of the points in the ith cluster, and

$$\|\mathbf{x}\|^2 = \sum_{j=1}^{d} x_j^2,$$

if $\mathbf{x} = (x_1, x_2, \ldots, x_d)$.

In class, we said that in general the k-means problem is NP-hard. However, for $d = 1$ the problem is polynomial. Design an algorithm that solves the k-means problem in time polynomial in the number of points n and the number of clusters k, for $d = 1$.

(**Hint:** Can you solve the problem for k clusters if you assume that you can solve it for fewer than k clusters?)

2. We are given a set P of n points in $\mathbb{R}$. For simplicity, assume that $\mu(P) = 0$, that is, $\sum_{x \in P} x = 0$. Let $\|P\|^2 = \sum_{x \in P} x^2$ be the optimal 1-means cost. Show that by adding carefully $O(1/\epsilon)$ centers, we can make the k-means cost at most $\epsilon \cdot \sum_{x \in P} x^2$.

Hint: First show that by adding 2 centers at locations $-\ell$ and ℓ, for an appropriate value of ℓ, the cost decreases by a factor of 3/4.