Data Mining

Homework 3

Due: 28/11/2021, 23:59

Instructions

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

The first homework has to be done by each **person individually**.

Handing in: You must hand in the homeworks by the due date and time by an email to Andrea (`mastropietro@diag.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 Andrea.

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

For any questions on the homework, clarifications, and so on, contact Andrea (`mastropietro@diag.uniroma1.it`).

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

Problem 1. For this exercise you will perform some clustering and use some feature engineering techniques. The lecture by Pablo Duboue[1] gives us a lot of interesting insights on the topic of feature engineering. It is a procedure that is sometimes underestimated but that, as you saw, can be really game changing in machine learning applications; the correct way to preprocess and model the features can increase the accuracy of many ML models. For this exercise we will use the Steam Reviews Dataset 2021 `https://www.kaggle.com/najzeko/steam-reviews-2021`. It presents different kinds of features, a fact that strongly suggests us to use some engineering techniques to better work with such data. What you have to do is the following:

1. Choose one clustering algorithm (k-means++, DBSCAN, etc.) to cluster the samples in the dataset **without** any feature engineering technique applied, just on raw data. You can use library implementations. Use the elbow method (`https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/`) to find the most suitable number of clusters. You are also free to use a different method to find the optimal number of clusters, if you want (like Silhouette, etc).

2. Inspect and understand your data and choose the best feature engineering technique (or techniques) that you would apply. Do whatever you think may be necessary (normalization,

[1] `http://duboue.net`

scaling, one-hot encoding and more). After that, perform the clustering again using the same clustering method used with the raw data.

Do you see any differences in the clusters obtained? What about running times? Make a plot of the clusters in both cases and comment what you see. Did your feature engineering help in creating more distinct clusters? Did it help with the shape of the elbow curve (or with the different method chosen)? Write a very short report (max 3 pages) in which you describe the clustering algorithm employed and the feature engineering techniques used (and why you think they were the proper techniques to use) along with the plots, comments and any observation on the results obtained that you think to be important. Hand in the report along with the code and instructions to run it.

Note. The dataset as is, is large. You can start by working on a subset of the dataset; 1GB of data should be reasonable.

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.