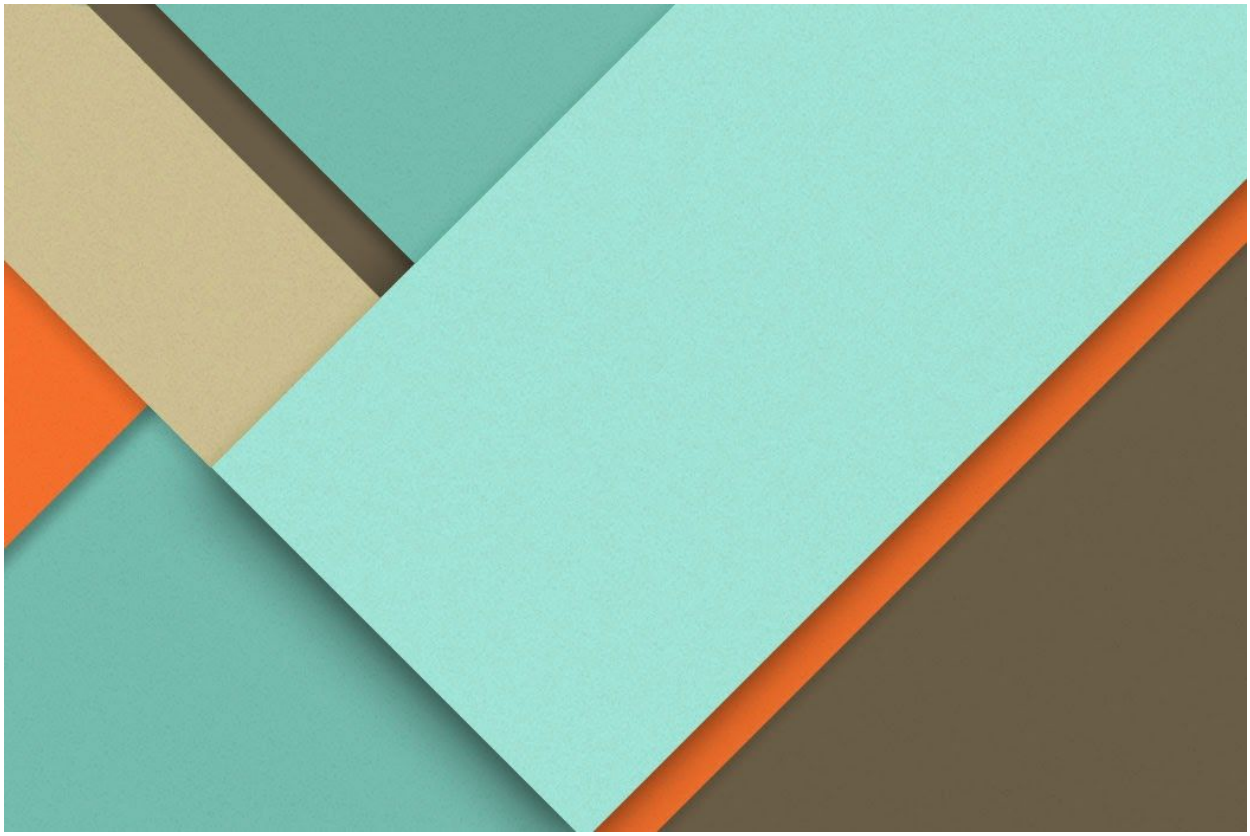


PRML Mid Sem Exam

27.10.2020

Vipul Yuvraj Nikam
180010041
3rd year, CSE

Question 1:

a. Let

$$f_1(x_1, x_2) = x_{21} + x_{22} + 2x_1 + 2x_2, \quad -10 \le x_1 \le 10, -10 \le x_2 \le 10$$

$$f_2(x_1, x_2) = x_1 \sin(x_1) + x_2 \sin(x_2), \quad -10 \le x_1 \le 10, -10 \le x_2 \le 10$$

For both the functions:

- Draw the surface using appropriate libraries and comment on the type of surface (convex/ non-convex).
- Find the minimum value of x_1 and x_2 analytically and verify the same from the plot.
- Use gradient descent and normal equation method, and obtain the minimum value of x_1 and x_2 . From the observation clearly illustrate the pros. and cons. of both techniques.
- Using various learning rates and initialization tabulate the results and justify your observation using appropriate theoretical illustration.

b. What is the difference between regression and k-means clustering tasks.

Based on the cost function of both the techniques, comment on the convergence (i.e will they reach the global minima). Justify your solution analytically (using appropriate mathematical equations or intuition).

- Take an appropriate example and illustrate the same (coding is expected).
- If you think that any of the methods will not reach the global minima, propose a technique that may help (to some extent) to reach the global minima. Justify your proposed method by extending your earlier illustration.?

Answers:

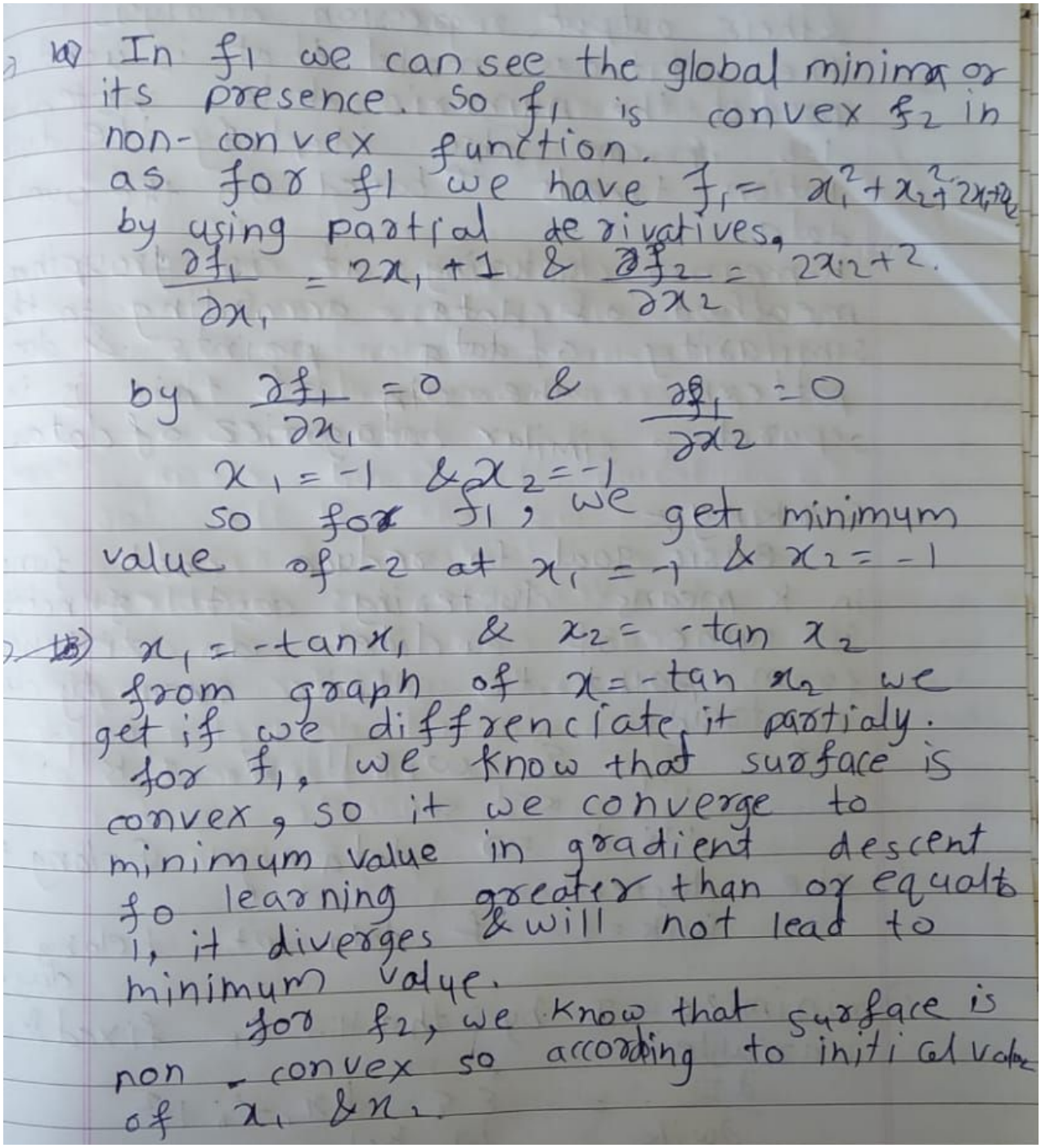
I. Code:

Click here :
https://colab.research.google.com/drive/13aTif8VL6M_8b3_vkcSC57GP8hDd1GqD?usp=sharing

II. Graphs & Plots:

All graphs are attached above in the code file.

III. Explanation:



a) In f_1 we can see the global minima or its presence. So f_1 is convex f_2 in non-convex function.
as for f1 we have $f_1 = x_1^2 + x_2^2 + 2x_1 + 2$
by using partial derivatives,
$\frac{\partial f_1}{\partial x_1} = 2x_1 + 1$ & $\frac{\partial f_2}{\partial x_2} = 2x_2 + 2$.

by $\frac{\partial f_1}{\partial x_1} = 0$ & $\frac{\partial f_1}{\partial x_2} = 0$

$x_1 = -1$ & $x_2 = -1$

so for f_1, we get minimum value of -2 at $x_1 = -1$ & $x_2 = -1$

b) $x_1 = -\tan x_1$ & $x_2 = -\tan x_2$

from graph of $x = -\tan x_2$ we get if we differenciate it partialy.
for f_1, we know that surface is convex, so it we converge to minimum value in gradient descent fo learning greater than or equalto 1, it diverges & will not lead to minimum value.

for f_2, we know that surface is non-convex so according to initial value of x_1 & x_2.

⇒ 16) Regression: it predicts the contineous value & their output. regression analysis is the statical as model that is used to predict the numeric data instead of leb it can also identify the distribution trends to based on the available data.

K-means clustering :- it is grouping meathod of data according to the similarity of data points & data pattern. the aim of this is to separate similar catagories of data.

working :-

Basic goal to reduce cost function in K means clustering cost function is same of euclide an distance from point to their mean by clus controid.

$$J = \sum_{i=1}^{} \sum_{k=1}^{} w_{i,k} \| x_i - c_k \|^2$$

$w_{i,k} = 0$ data pt belong to the cluster.

$w_{ik} = 1$ if data pt belong to k clust

i) minimize I by thanHin c_k fixed & w_{ik} variable:

$$\frac{\partial I}{\partial w_{ik}} = \sum_{i=1}^{m} \sum_{k=1} \| x_i - c_i \|^2$$

reducing this g will get:

$$\begin{cases} 1 & ; \text{if } k = \text{argonil} \| x_i - c_k \| \\ 0 & \text{otherwise.} \end{cases}$$

2) Minimize by treating w_{ik} constest & c_k variable.

$$\frac{\partial I}{\partial c_k} = 2 \frac{k}{2} \; w_{ik} \| x_i - c_k \|$$

we will get the following by reducing

$$c_k = \frac{\sum_{i=1}^{m} w_{ik} x_i}{\sum_{i=1} w_{i\,k}}$$

The above two step will be carized out iterating unit we get the optimal segration of data point.

Question 2:

Consider the credit card fraud detection task (downloaded the data-set from: kaggle- link).

- Follow these steps to pre-process the data and obtain its subset:
    1. Export the data in .csv file using any library of your choice.
    2. As the dataset is imbalanced (i.e less no. of fraud examples compared to genuine), choose randomly the same number of genuine examples as fraud examples and obtain a balanced dataset .
    3. Remove the outliers from the selected examples (use threshold of 2.5 IQR).
    4. Normalize the data (features should have zero mean and unit standard devia-tion).
    5. Use this data for further processing.
- Use t-SNE plot (refer: t-SNE link) to visualize the subset of data in 2D. Intuitively comment which clustering algorithm (k-means, GMM, agglomerative hierarchical clustering, DBSCAN) will work better and provide justification for the same using your theoretical knowledge.
- Propose and clearly explain an evaluation measure to evaluate the clustering algorithms. Using a piece of code, prove your intuition of clustering algorithm selection using the proposed evaluation measure (use the same subset of data obtained in the first part).
- Take the whole (imbalanced) data, do the data pre-processing task (out-lier removal, data normalization), perform clustering using all the above mentioned techniques. If degradation in performance is observed, comment on the possible causes for it. Propose a possible solution to overcome this issue (without reducing the size of the dataset). Justify the same through a code.
- Split the whole data in a training and testing set (90%, 10%, perform class specific division as the data is imbalanced). Use any one of the above mentioned clustering techniques and perform classification tasks (using a piece of python code). In the report, clearly mention the steps involved in training and testing, and the performance measure with appropriate mathematical equations.?

Answers:

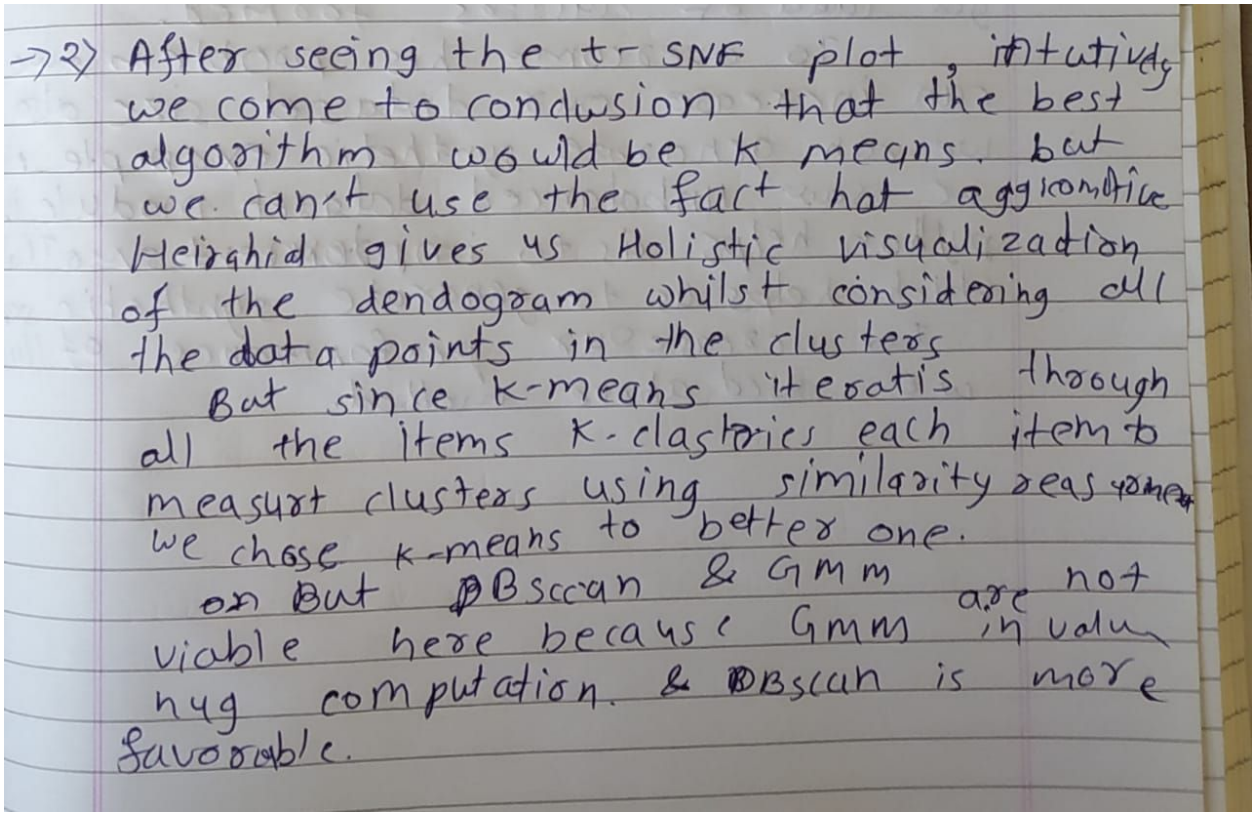
I. Code:

Click here :
https://colab.research.google.com/drive/1HxdPtYtILWWgbZIfoahkPzkO-LV98s4f?usp=sharing

II. Graphs & Plots:

All graphs are attached above in the code file.

III. Explanation:



$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i} d(i,j)$ $\quad b_i = \min \left(\frac{1}{|C_I|} \sum d_{ij} \right)$

$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$ where $C_{(i)}$ is cluster index.

$d(i,j)$ ith data.

$|C_i|$ no. of data pt.

so, from the plot we can see that DBSCAN & GMM have performed poorly than K-means & agglometive ~~the~~ compair, but DBSCAN was better than GMM. when whole data set was considered for clustering, the fraud clean wasn't visible. so we built all the algo. to get idea of clusters formed.

To overcome this problem, we can increase our dataset size also we can genarate synthetic example. we saw unbalanced dataset module in python helps us to implement SMOTE algo which helps us to genarate synthetic sample.

To increase our performance of clusters we used above method.