Project 3 Tasks

The whole submission consists of two parts:

1. Task I : District Clustering based on Human Development [33 marks]
2. Task II: PCA Implementation using numpy [7 marks]

So, the total is 40 marks.

Task 1 Description

Using the dataset given below in Set 2, cluster the 75 districts (it was 75 when that data was collected) based on their "human development" as given by education (literacy), life expectancy, and income.

The context is that we need to figure out which districts are "similar" so that it can guide human development related policy. We could just find the Human Development Index (HDI) and sort accordingly but with this work, we want a more nuanced approach.

Algorithms to Try

Please use the following clustering methods on the data:

1. K-means
2. DBSCAN
3. Gaussian Mixture Model

Report Format:

Note: Also include the link to code.

1. **Data exploration and analysis**
   Data analysis and visualization. Also mention findings.
2. **Feature Selection and preprocessing:**
   Add commentary on what features are being used and the reason behind the preprocessing done
3. **K-Means | Approach taken and findings**
4. **DBSCAN | Approach taken and findings**
5. GMM | Approach taken and findings
6. Final Recommendation of clusters

For details on expectation, please go over the Grading Criteria.
Also, this work covers a lot of what we expect.

Datasets:

Set 1: Meat and eggs | Used in class

All agriculture

Meat production

Egg production

Set 2: Human Development Index | To be used for Project 3

S2.1: Life Expectancy and income

S2.2: Literacy rate by gender

Possible extra data:
Education level for people 5 years and older

Population by federal units

Number of federal units

Resources

Reference
http://jasontdean.com/python/seattleCrime.html

Useful Links

https://www.quora.com/Cluster-Analysis/Can-the-choice-of-the-measurement-units-of-the-features-impact-the-clustering/answer/Franck-Dernoncourt

How to understand the drawbacks of K-means

How to Determine the Optimal K for K-Means? - Analytics Vidhya

What is clustering and why is it hard? · Its Neuronal

In Depth: k-Means Clustering | Python Data Science Handbook

How to compare dbscan clusters / choose epsilon parameter

http://web.iitd.ac.in/~bspanda/dbscan%20clustering.pdf

Task 1 | Grading Criteria

* HDI Criteria: Literacy, Income, Life Expectancy

Total Marks: 33

Header	**Points**	**Description**
Data Usage	4	1 - Has made use of only two of the HDI criteria 3 - Has made use of all three HDI criteria by merging ***S2.1*** and ***S2.2*** 4 - Has ingested in data apart from ***S2.1*** and ***S2.2***
Data Preparation	2	1 - Has cleaned the data to have one entry for each district +1 Also has commentary along the way to explain the steps done and why
Data Exploration and Analysis	4	1 - Has performed minimal data exploration 2 - Has performed good amount of data exploration +1 Also has commentary on the findings +1 Also has connected it to how it has influenced further work below (e.g: were some columns discarded based on the analysis?)
Data Preprocessing	3	+ 1 Has been deliberate about the columns/features chosen for clustering with the reasoning specified + 1 Has performed feature standardization / normalization +1 Also has provided justification of the choice and why it needs to be done
K-means	6	1 - Has used K-means to cluster the data + 1 Has used elbow-method to try to choose the number of k + 1 Also has looked at the silhouette score to guide the choice of k + 1 Has justification on final value of k chosen (could be in contrast to the suggestion from the scores if the reasoning is good) + 1 Has used PCA to visualize the clusters formed + 1 Has commentary on what the clustering has said about the data; could be based on just the final clustering achieved or the process as a whole
DBSCAN	6	1 - Has used DBSCAN to cluster the data + 1 Has used k-neighbors to try to choose the value for epsilon + 1 Has justification on final value of epsilon chosen

		(could be in contrast to the suggestion from the scores if the reasoning is good) + 1 Has justification on the final value of `min_samples` chosen + 1 Has used PCA to visualize the clusters formed + 1 Has commentary on what the clustering has said about the data; could be based on just the final clustering achieved or the process as a whole
GMM	3	1 - Has used GMM on the data + 1 - Has justified choice of number of components + 1 - Has commentary on what the clustering has said about the data; could be based on just the final clustering achieved or the process as a whole
Final Clustering	2	1 - Has provided a final clustering based on all the work +1 Also has justification on why the choice was made
Report	3	1 - Slapdash work: E.g. no attention paid to formatting 2 - Well presented report: E.g: good headers and subheaders 3 - Professionally Done report. E.g; Also makes use of good practices like highlighting performance of the best model on the comparison table. Plots well titled and labeled. Overall, the report is easy to follow and consume.

Task 2 | Description

Implement PCA using numpy / scipy. *Note: You don't need to submit a report for this. Only the notebook.*

Grading Criteria

Header	Points	Description
Sorting "eigen_values and eigen_vectors"	1	+0.5 Sort the eigen value in decreasing order +0.5 Sort the eigenvector in the decreasing order of eigenvalue.
Taking the first components	0.5	+ 0.5 Take the first num_of_component components from the eigen_values and eigen_vectors
Dimensionality reduction	0.5	Perform the dimensionality reduction to get lower dimension data by multiplying data matrix with first num_of_component components eigen_vectors matrix
Visualization of the lower dimension data	2.5	+1 Use Scatter plot where 1st Component in X-axis and 2nd Component in Y-axis +1 Use target information to identify which data point belongs to which digit to do the color coding and show the legend in the plot. +0.5 Put title in the plot and resize the plot according to the need
num_of_component according to the variance explained	2.5	+0.5 Calculate the variance explained by each eigenvalue +0.5 Calculate the total variance explained by the first K eigen value +0.25 Do line plot where K value in X-axis, total_variance[K] value in Y-axis using parameter marker='o', linestyle='--', color='b' +0.25 Put Proper x-label, y-label and title in the graph +0.25 Show vertical grid line in the graph +0.25 Plot one horizontal red color line at the chosen threshold level. +0.5 Conclusion that you get from the graph