UNSUPERVISED LEARNING

TOTAL MARKS:70

DATASET:

Note: For this exam two datasets are used. Q.1 and 2 should be answered for both the datasets. The first dataset dermatology.csv should be used for answering questions 3 second dataset should be used for answering question 5 and 6. Question no 6. should summarize all the findings for both the datasets.

This database (dermatology.csv)contains 34 attributes, 33 of which are linear valued and one of them is nominal.

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

Attribute Information:

Clinical Attributes: (take values 0, 1, 2, 3, unless otherwise indicated) 1:
erythema
2: scaling
3: definite borders
4: itching
5: koebner phenomenon
6: polygonal papules
7: follicular papules
8: oral mucosal involvement
9: knee and elbow involvement

10: scalp involvement
11: family history, (0 or 1)
34: Age (linear)

Histopathological Attributes: (take values 0, 1, 2, 3)
12: melanin incontinence
13: eosinophils in the infiltrate
14: PNL infiltrate
15: fibrosis of the papillary dermis
16: exocytosis
17: acanthosis
18: hyperkeratosis
19: parakeratosis
20: clubbing of the rete ridges
21: elongation of the rete ridges
22: thinning of the suprapapillary epidermis
23: spongiform pustule
24: munro microabcess
25: focal hypergranulosis
26: disappearance of the granular layer
27: vacuolisation and damage of basal layer
28: spongiosis
29: saw-tooth appearance of retes
30: follicular horn plug
31: perifollicular parakeratosis
32: inflammatory monoluclear inflitrate
33: band-like infiltrate
35:class label

1. Data Understanding (5 marks)

a. Read the dataset (tab, csv, xls, txt, inbuilt dataset). What are the number of rows and no. of cols & types of variables (continuous, categorical etc.)? (1 MARK)

b. Calculate five-point summary for numerical variables (1 MARK)

c. Summarize observations for categorical variables – no. of categories, % observations in each category. (1 MARK)

d. Generate the covariance and correlation tables for the data (1 MARK)

e. Create Visualization plots to find the relationship amongst the variables. (1 MARK)

2. Data Preparation (5 marks)

a. Scale / Transform/ clean the data so that it is suitable for model building.

3. Dimensionality Reduction (15 marks)

a. How will you decide when to apply PCA based on the correlation? (2 marks)

b. Apply PCA on the above dataset and determine the number of PCA components to be used so that 90% of the variance in data is explained by the same. (7 marks)

c. Build a data frame with the principal components and check if multi-collinearity still exists. (2 marks).

d. Visualize the spread of data across PCA components. (2 marks)

e. Check for outliers in the PCA data and treat the same. (2 marks)

4. Clustering: Use PCA dimensions to cluster the data. Apply K-means and Agglomerative clustering. (30 Marks)

Some pointers which would help you, but don't be limited by these

a. Find the optimal K Value. (3 marks)

b. Apply Clustering and visualize the spread of data (20 marks)

c. Evaluate the clusters formed using appropriate metrics (inertia, silhouette score) to support the model built and compare both the models. (5 marks)

d. Using best attributes based on the relationship between them, plot the clusters. (2 marks)

5. Use the cluster labels from the best method above and interpret the clusters formed. (5 marks)

6. Summarize as follows (10 marks)

a. Summarize the overall fit of the model. Compare all the clustering models built and list down the measures to prove that it is a good model.

b. Write down a business interpretation/explanation of the model. (List the countries to be focused on).

c. What are the key risks to your results and interpretation?