Machine Learning Homework (3)

Clustering measures (20 points)

- Clustering Iris data with Kmeans and DBSCAN with the normalization method you select, then code the following clustering measures and compare two algorithms' performance differences as well as parameter setting in DBSCAN[1]
  - adjusted_mutual_info
  - homogeneity
  - completeness
  - v_measure

[1] You need to code the clustering measures from scratch rather than call sklearn related functions. But you can use the functions to validate your coding.

Credit Risk Analytics via Machine learning (20 points)

- We have the following data set for credit ranking for 12 different industry sections (it is a simulated data):
  - *credit_sim_ data.csv*, where the first **1540** samples (rows) are labeled as 'good credit' (label type: '1'), i.e., whose credit rankings are 'AAA', 'AA', or 'A'
  - and the remaining **130** samples are labeled as ' bad credit', (label type: '0') whose credit ranks are 'CCC'.
- There are six variables (columns) in this data set:

```
variable 1: Working capital / Total Assets (WC_TA)
variable 2: Retained Earnings / Total Assets (RE_TA)
variable 3: Earnings Before Interests and Taxes / Total Assets (EBIT_TA)
variable 4: Market Value of Equity / Book Value of Total Debt (MVE_BVTD)
variable 5: Sales / Total Assets (S_TA)
variable 6: Industry sector labels from 1-12
```

- Complete the following problems
  - Use at least 4 normalization methods to normalize data.
  - Conduct 80% and 20% training-test split and k-fold (k=10) cross validation for the data and use the following prediction to conduct classifications and compare their results
    - k-NN with the following distances: Euclidean, "cosine', 'correlation', 'chebyshev', and 'seuclidean'
  - Use confusionmatrixdisplay in sklearn.metrics to visualize the confusion matrix you get from each classification cases.
  - Calculate all classification measures including d-index to evaluate learning results and draw your conclusion.

Query PM 2.5 (Air Quality Index) (30 points)

PM2.5 refers to dangerous atmospheric particulate matter (PM) that have a diameter less than 2.5 micrometers. It acts as an important measure for air population in big cities. Seniors and children are very sensitive to PM2.5.

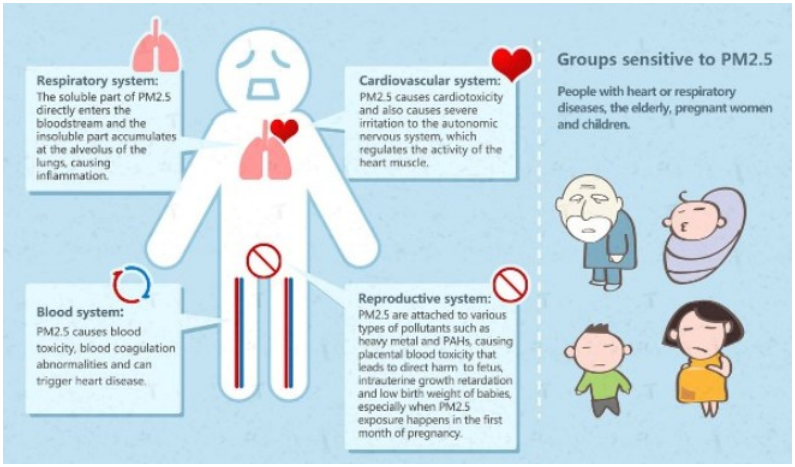

Fig 1. PM2.5 and its sensitive groups.

A research group at TsingHua University collected hourly PM2.5 data in Beijing from 2010 to 2014 and obtained total 48,324 hourly samples. They used the following features in their data collection.

year: year of data in this row

No: row number

month: month of data in this row day:

day of data in this row

hour: hour of data in this row pm2.5: PM2.5 concentration (ug/m^3)

DEWP: Dew Point (â„f)

TEMP: Temperature (â„f)

PRES: Pressure (hPa)

cbwd: Combined wind direction

Iws: Cumulated wind speed (m/s)

Is: Cumulated hours of snow

Ir: Cumulated hours of rain

Finish the following problems

1. Clean the data by removing all the rows with missing information and obtain a new file called clean_pm2.5_data.csv
2. Clustering PM 2.5 data in 2010, 2011, 2012, 2013, 2014 with bisecting KMmeans respectively. Each dataset should contain have 3 clusters (low, medium and high PM 2.5). Discuss your results.
   (a) Note: you need drop some non-numerical features or change them to numerical ones
   (b) Features 'Is' and Ir' should be dropped
3. Use all samples in 2010–2013 as training data to train a k-NN and use all samples in 2014 as test data to verify your k-NN learning machine's performance. You should make any samples with PM2.5 >=100 as ('high') and others as low.
   - k-NN at least with the following distances: Euclidean, "cosine', 'correlation', 'chebyshev', and 'seuclidean'

What should you turn in?

- 1. A folder that contains
  - Detailed answers
  - 30-page PPT
  - data if available
    - source files