CS 584-04: Machine Learning

Fall 2018 Assignment 4

Diego Martin Crespo A20432558

Question 1 (60 points)

Diabetes is a true public health problem. According to a 2016 report by the Illinois Department of Public Health, about 1.3 million people in Illinois, or 12.8% of the population have diabetes. Another estimated 341,000 people have diabetes but don't know about it. In Chicago, 25.6% of the population have been hospitalized due to diabetes in 2011. However, health activists have long suspected that there are clusters of populations in Chicago which have more serious diabetes health problems.

You are asked to analyze the ChicagoDiabetes.csv file to identify clusters of diabetes population. This CSV file is extracted from the data which can be downloaded from the Chicago Data Portal https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Diabetes-hospitalizations/vekt-28b5. The ChicagoDiabetes.csv contains the annual number of hospital discharges and the crude hospitalization rates, for the years 2000 to 2011, by the 46 communities of Chicago. The crude hospitalization rate is the number of hospital discharges in a community divided by the total population for the community. The crude rates are expressed per 10,000 residents.

You will perform the following analyses in your research.

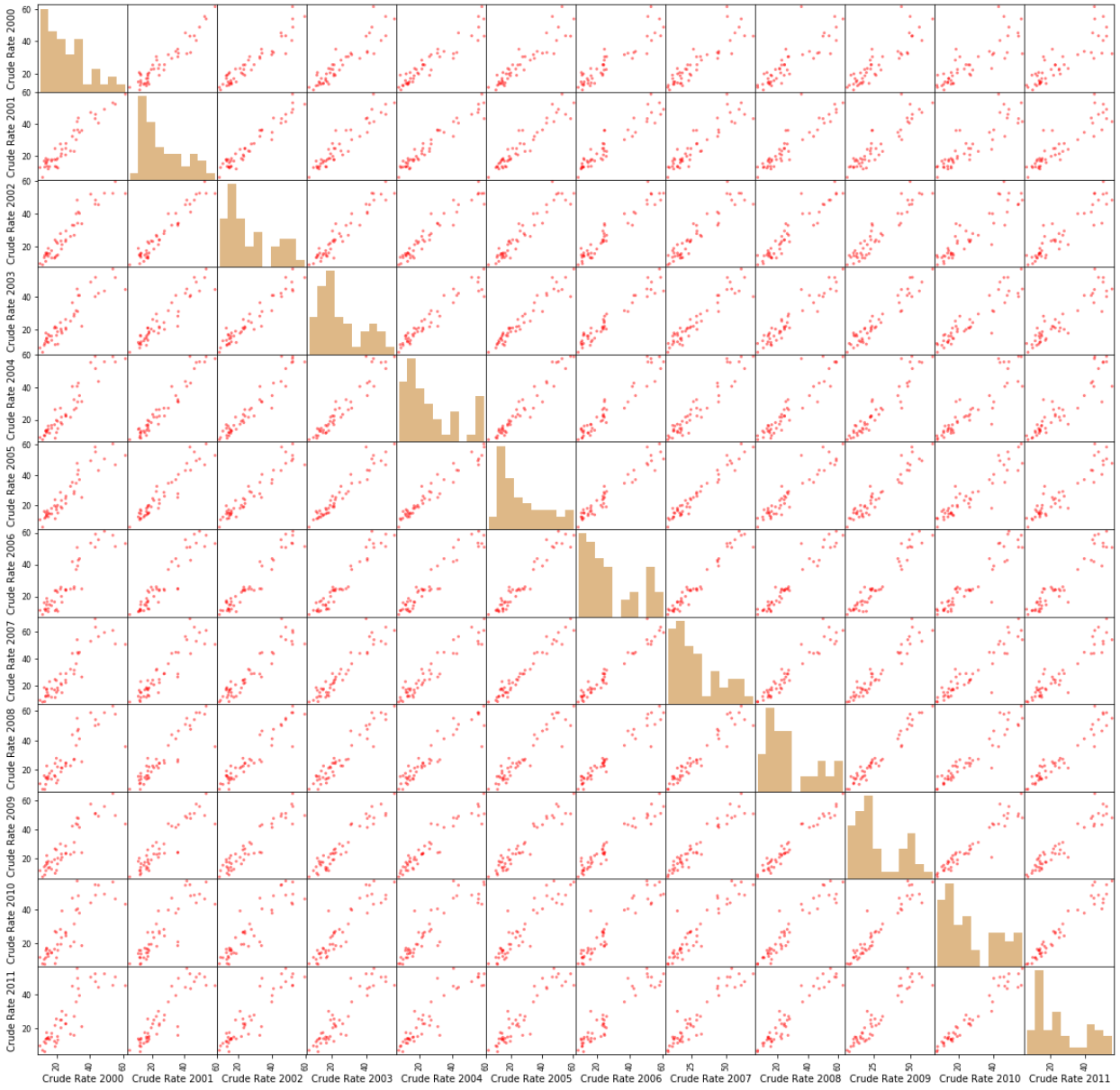
1. You use only the 12 crude hospitalization rates (i.e., Crude Rate 2000 to Crude Rate 2011) in a Principal Component analysis.
2. You use the first two principal components in a Clustering analysis. You search the number of clusters from 2 to 15.
3. You calculate the annual total population and the annual number of hospital discharges in each cluster for each year, then calculate crude hospitalization rate in each cluster for each year.
4. You plot the crude hospitalization rates in each cluster against the years. You also plot the Chicago's annual crude hospitalization rates against the years as the reference curve.

After you have completed your analyses, please answer the following questions.

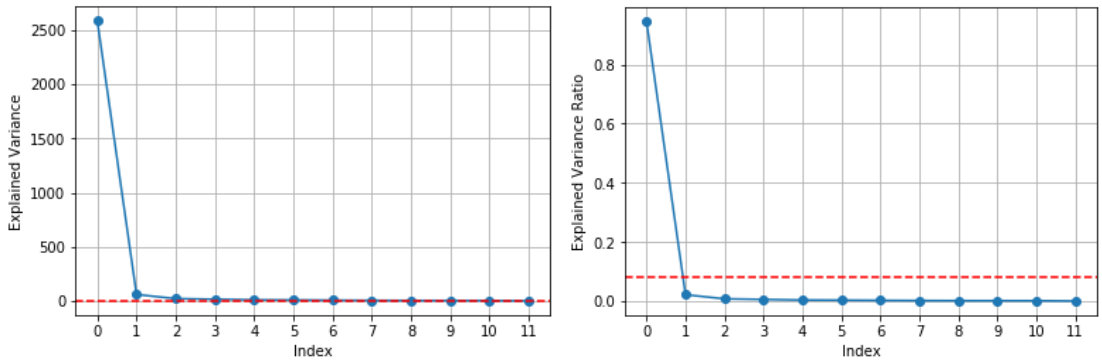
a) (5 points). How many observations and variables did you use in your Principal Component analysis?

- 12 variables
- 46 observations

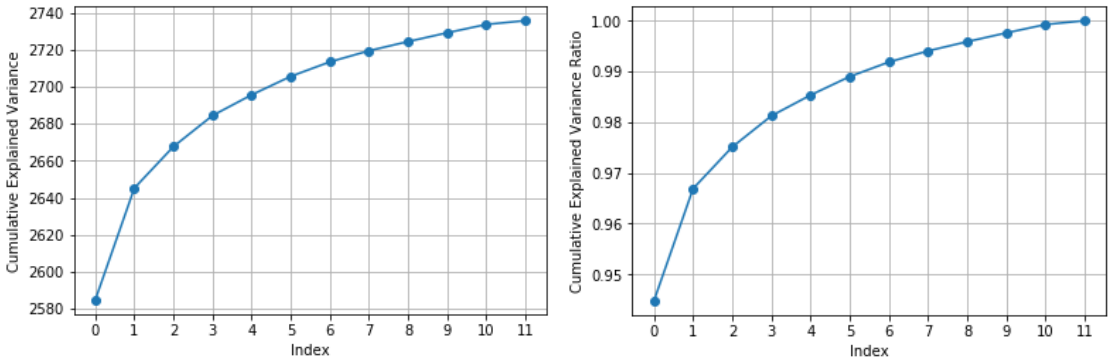
b) (5 points). Generate the scatter plot matrix for the variables. Put the histograms on the diagonal.



c) (5 points). Plot the Explained Variances against their indices. Add a horizontal reference line whose value is the reciprocal of the number of variables. Label the axes and add grid lines to the axes.



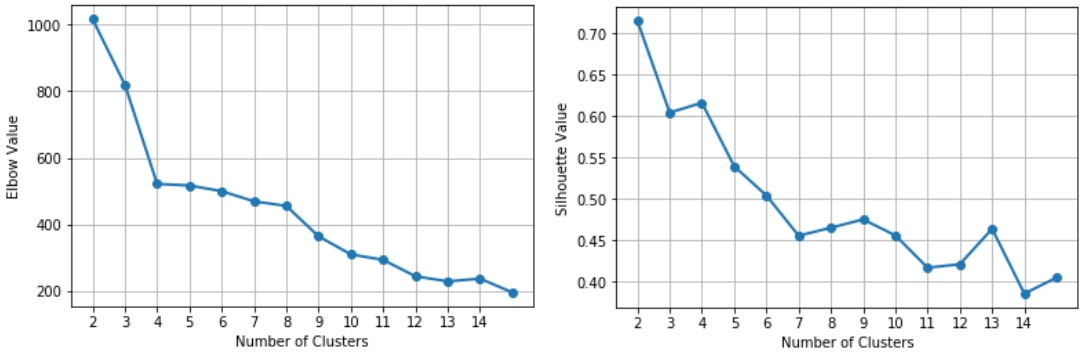
d) (5 points). Plot the Cumulative Sum of the Explained Variances against their indices. Label the axes and add grid lines to the axes.



e) (5 points). What percentage of the total variance is explained by the first two principal components?

A 96.6902% ((2645.16/2735.71) x 100) is explained by the first two components

f) (5 points). Plot the Elbow and the Silhouette charts against the number of clusters.



g) (5 points). What is the number of clusters that you will choose based on the charts in f)?

The number of chosen clusters is 4, as it is at the “elbow” in the Elbow graph and also has one of the highest values in the silhouette graph.

h) (5 points). How many communities are in each cluster?

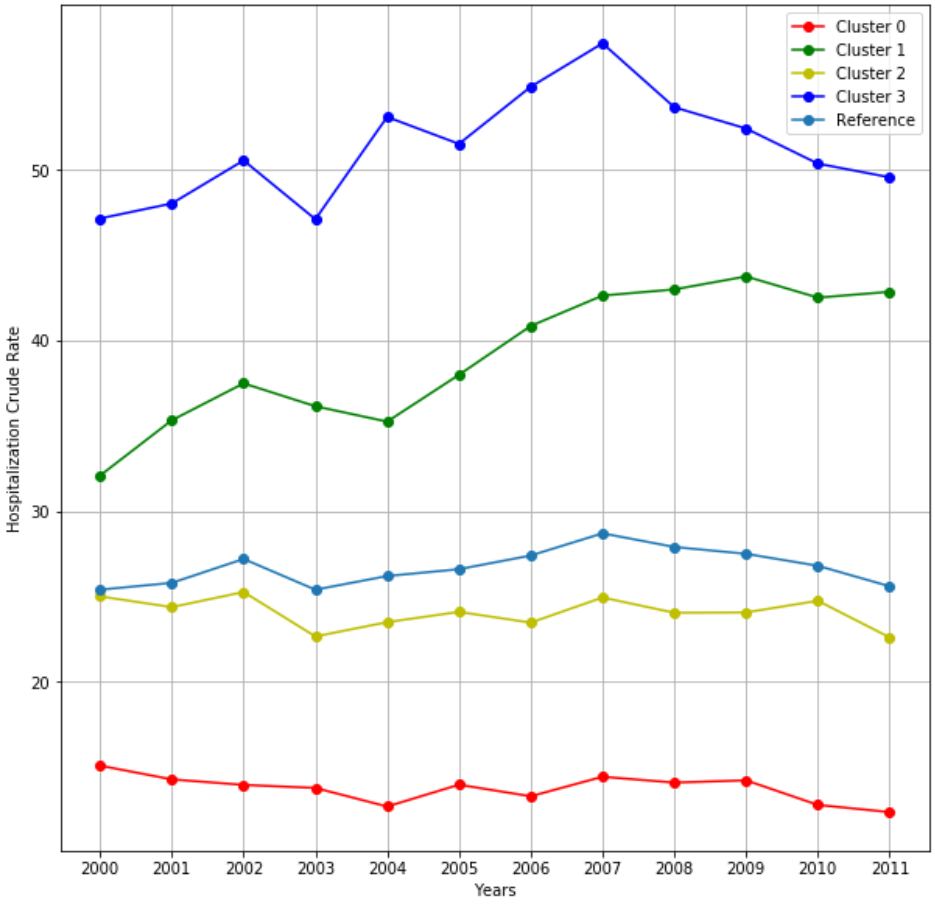
	CLUSTER 0	CLUSTER 1	CLUSTER 2	CLUSTER 3
Nº Communities	18	5	14	9

i) (5 points). List the names of the communities in each cluster.

	CLUSTER 0	CLUSTER 1	CLUSTER 2	CLUSTER 3
Communities	'Downtown', 'West Loop', 'Near North Side', 'Lake View', 'Lincoln Park', 'Avondale', 'Albany Park', 'Jefferson Park', 'Edison Park', 'Archer Heights', 'Dunning', 'Portage Park', 'West Ridge', 'Edgebrook', 'Mount Greenwood', 'Norwood Park', 'Belmont Harbor', 'North Park'	'Near West Side', 'South Chicago', 'Woodlawn', 'Beverly', 'West Humboldt Park'	'Lower West Side', 'New City', 'Hyde Park', 'Chinatown', 'West Town', 'South Lawndale', 'Rogers Park', 'West Lawn', 'Garfield Ridge', 'Belmont Gardens', 'Edgewater', 'Bucktown', 'Ashburn', 'Edgewater Glen'	'Chatham', 'Auburn Gresham', 'Englewood', 'West Garfield Park', 'Roseland', 'West Englewood', 'Austin', ' South Shore', 'Kenwood'

j) (10 points). Plot the crude hospitalization rates in each cluster against the years. You also plot the Chicago's annual crude hospitalization rates (in the table below) against the years as the reference curve.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Rate	25.4	25.8	27.2	25.4	26.2	26.6	27.4	28.7	27.9	27.5	26.8	25.6



k) (5 points) Based on the graph in j), what will you conclude about the trend of crude hospitalization rate in each cluster relative to the Chicago's rates?

Both clusters 1 and 3 are above the reference rates from Chicago. Meanwhile clusters 0 and 2 are below the reference curve. The cluster 2 have a similar trend to the reference rates of Chicago.

Question 2 (40 points)

In 2014, Allstate provided the data on Kaggle.com for the Allstate Purchase Prediction Challenge which is open. The data contain transaction history for customers that ended up purchasing a policy. For each Customer ID, you are given their quote history and the coverage options they purchased.

The data is available on the Blackboard as Purchase_Likelihood.csv. It contains 665,249 observations on 97,009 unique Customer ID. You will use the **empirical** Naïve Bayes model to predict purchase likelihood of coverage A using three predictors. The target variable is **A** which have these categories 0, 1, and 2. The nominal predictors are (categories are inside the parentheses):

1. **group_size**. How many people will be covered under the policy (1, 2, 3 or 4)?
2. **homeowner**. Whether the customer owns a home or not (0=no, 1=yes)
3. **married_couple**. Does the customer group contain a married couple (0=no, 1=yes)

After you have built the empirical Naïve Bayes model, then please answer the following questions.

a) (5 points) What are the Class Probabilities?

Class A	Nº of class A	Class Probability
0	143691	0.2159958
1	426067	0.6404624
2	95491	0.1435417

b) (5 points) When group_size = 1, homeowner = 0, and married_couple = 0, what are the predicted probabilities Pr(A = 0), Pr(A = 1), and Pr(A = 2)?
- Pr(A = 0)= 0.22703676
- Pr(A = 1)= 0.62759396
- Pr(A = 2)= 0.14536928

c) (5 points) When group_size = 2, homeowner = 1, and married_couple = 1, what are the predicted probabilities Pr(A = 0), Pr(A = 1), and Pr(A = 2)?
- Pr(A = 0)= 0.20408045
- Pr(A = 1)= 0.65112388
- Pr(A = 2)= 0.14479567

d) (5 points) When group_size = 3, homeowner = 1, and married_couple = 1, what are the predicted probabilities Pr(A = 0), Pr(A = 1), and Pr(A = 2)?
- Pr(A = 0)= 0.21468565
- Pr(A = 1)= 0.6385567
- Pr(A = 2)= 0.14675765

e) (5 points) When group_size = 4, homeowner = 0, and married_couple = 0, what are the predicted probabilities Pr(A = 0), Pr(A = 1), and Pr(A = 2)?
- Pr(A = 0)= 0.26230538
- Pr(A = 0)= 0.58747922
- Pr(A = 0)= 0.1502154

f) (10 points) What are the values of the predictors group_size, homeowner, and married_couple such that Prob(A = 1) attains its maximum?

Prob(A = 1)_max= 0.6634093568143489

- Group_size= 1
- Homeowner= 1
- Married_couple=1

g) (5 points) For the values of group_size, homeowner, and married_couple, what are the predicted probabilities Pr(A = 0), Pr(A = 1), and Pr(A = 2)?

Group_size	Homeowner	Married_couple	Pr(A = 0)	Pr(A = 1)	Pr(A = 2)
1	1	1	0.19384456	0.66340936	0.14274608
1	0	1	0.21439256	0.63746296	0.14814447
1	1	0	0.20558784	0.65412798	0.14028418
1	0	0	0.22703676	0.62759396	0.14536928
2	1	1	0.20408045	0.65112388	0.14479567
2	0	1	0.22534324	0.62463169	0.15002508
2	1	0	0.21628018	0.64152898	0.14219084
2	0	0	0.23844031	0.61446407	0.14709562
3	1	1	0.21468565	0.6385567	0.14675765
3	0	1	0.2366541	0.61154408	0.15180182
3	1	0	0.22734114	0.62865421	0.14400465
3	0	0	0.25019947	0.60108716	0.14871338
4	1	1	0.22565685	0.62571883	0.14862432
4	0	1	0.24831841	0.59821444	0.15346716
4	1	0	0.23876549	0.61551636	0.14571815
4	0	0	0.26230538	0.58747922	0.1502154