CS 584: Machine Learning

Pradyot Mayank
A20405826
Spring 2019 Assignment 4

Question 1 (50 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 advocacy groups have long argued 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 of the community. The crude rates are expressed per 10,000 residents.

You are asked to perform the following analyses in your research.

1. Use only the non-missing values of the twelve crude hospitalization rates (i.e., Crude Rate 2000 to Crude Rate 2011) in a Principal Component analysis on the correlation matrix.

2. Select the major principal components in a Clustering analysis. Use this integer 20190405 for the random_state value in the KMeans function.

3. Search the number of clusters from two to ten.

4. 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.

5. Plot the crude hospitalization rates in each cluster against the years. Also, plot the Chicago's annual crude hospitalization rates against the years as the reference curve.

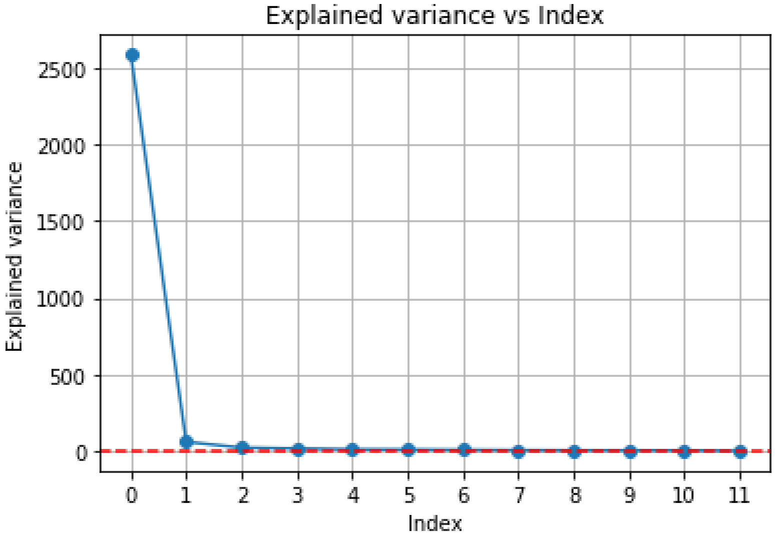
After you have completed your analyses, please answer the following questions.

a) (5 points). What is the maximum number of principal components that you can get?

The maximum number of Principal Components are 12

b) (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.

The required chart is given below with a horizontal reference line whose value is the reciprocal of the number of variables:



c) (5 points). Suppose I am required to explain at least 95% of the total variance, then which major (i.e., top *k*) principal components should I select?

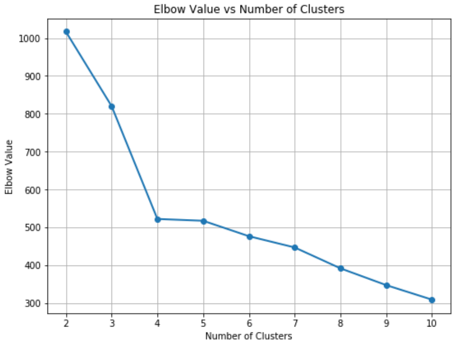
To explain at least 95% of total variance we need first two or top two Principal Components ie. PC1 and PC2 so k=2.

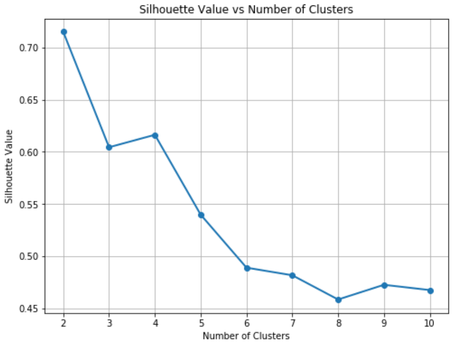
d) (5 points). What is the cumulative explained variance ratio accounted by the major principal components that you selected in c)?

The cumulative explained variance ratio accounted by the major principal components that we selected in (c) is 0.96690181.

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

The required the Elbow and the Silhouette charts are given below:





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

The number of clusters based on charts in e) are 4 as elbow is forming when number of clusters are 4 and second highest Silhouette value is attained when the number of clusters are 4.

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

Cluster 0= Downtown, Belmont Harbor, Norwood Park, Mount Greenwood, Edgebrook, West Ridge, Portage Park, Dunning, Archer Heights, Edison Park, North Park, Albany Park, Jefferson Park, Near North Side, West Loop, Lake View, Avondale, Lincoln Park.

Cluster 1= Austin, Chatham, South Shore, West Englewood, Auburn Gresham, Kenwood, Roseland, West Garfield Park, Englewood.

Cluster 2= Near West Side, Woodlawn, Beverly, South Chicago, West Humboldt Park.

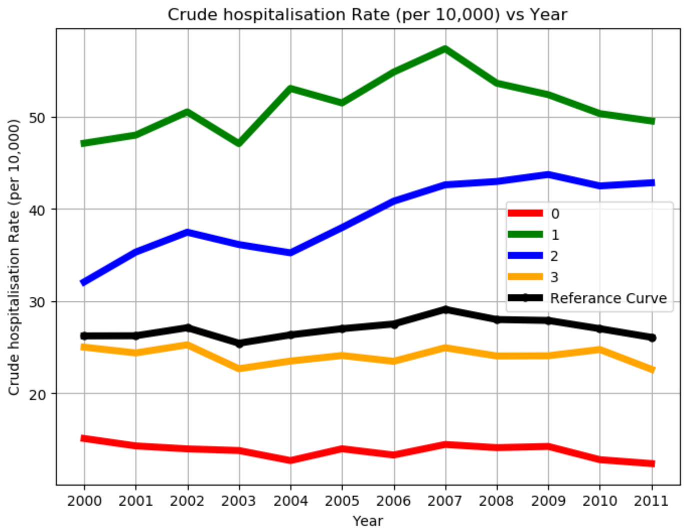
Cluster 3= Ashburn, New City, Bucktown, Lower West Side, Belmont Gardens, Garfield Ridge, Hyde Park, Chinatown, West Lawn, Rogers Park, South Lawndale, West Town, Edgewater, Edgewater Glen.

h) (5 points). What are Chicago's annual crude hospitalization rates from 2000 to 2011? Please present your answer in a table.

The Chicago's annual crude hospitalization rates from 2000 to 2011 is given below:

Years	Annual Crude Hospitalization rate
2000	26.22826
2001	26.24565
2002	27.11739
2003	25.43696
2004	26.33913
2005	27.01087
2006	27.51957
2007	29.10652
2008	28.00652
2009	27.90435
2010	27.01087
2011	26.07391

i) (5 points). Plot the crude hospitalization rates in each cluster against the years. You also plot the Chicago's annual crude hospitalization rates (in your answer in h) against the years as the reference curve.



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

Based on the curve above we can say that the crude rate of cluster 1 is the highest amongst all clusters. Cluster 0 and Cluster 3 are performing worse than the reference curve whereas cluster 2 and Cluster 1 are performing better than the Chicago crude rate reference curve.

Cluster 1 has the highest crude rate in 2007 and it decreases gradually till 2011.

Cluster 2 has the highest crude rate in 2009 and changes barely till 2011.

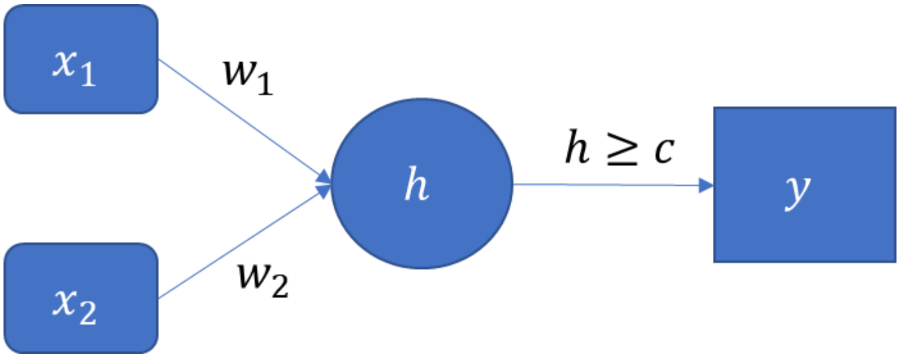
Cluster 0 and cluster 3 barely changes over the course of 12 years.

Question 2 (50 points)

Logical operators (i.e., AND, OR, XAND, etc.) are the building blocks of any computational device. Logical functions return only two possible values, TRUE or FALSE, based on the truth or false values of their input values. For example, the operator OR returns FALSE only when all the input values are FALSE. Otherwise, the operator OR returns TRUE. If we denote TRUE by 1 and FALSE by 0, then the logical OR function can be represented by the following table:

x_1	0	0	1	1
x_2	0	1	0	1
x_1 OR x_2	0	1	1	1

This function can be implemented by a perceptron with two binary inputs:



The activation functions for all the layers have this form: $y = \varphi(h) = 1$ if $h \geq c$. Otherwise, $\varphi(h) = 0$.

a) (15 points). If we restrict the values of the parameters w_1, w_2, and c to positive integers, then specify the lowest possible values for these parameters such that the perceptron can implement the logical OR function. You have to prove that your solution does work.

Since, the values of w1, w2 and c can only be positive integer values. So the lowest possible values of the parameters will be:

w1=1, w2=1 and c=1.

The activation function will be $\varphi(h)$= x1*w1+x2*w2 and to activate the function the value of c must be greater than or equal to 1.

Proof:

Case1: 0*1+0*1=0

Case2: 0*1+1*1=1

Case3: 1*1+0*1=1

Case4: 1*1+1*1=2

So, for the Case2, Case3 and Case4 the activation function of perceptron $y = \varphi(h)$ wiil be activated as its value is 2 which greater than or equal to c.

b) (15 points). If we restrict the values of the parameters w_1, w_2, and c to positive integers, then specify the lowest possible values for these parameters such that the perceptron can implement the logical AND function which can be represented by the following table. You have to prove that your solution does work.

x_1	0	0	1	1
x_2	0	1	0	1
x_1 AND x_2	0	0	0	1

Since, the values of w1, w2 and c can only be positive integer values. So the lowest possible values of the parameters will be:

w1=1, w2=1 and c=2.

The activation function will be $\varphi(h)$= x1*w1+x2*w2 and to activate the function the value of c must be greater than or equal to 2.

Proof:

Case1: 0*1+0*1=0

Case2: 0*1+1*1=1

Case3: 1*1+0*1=1

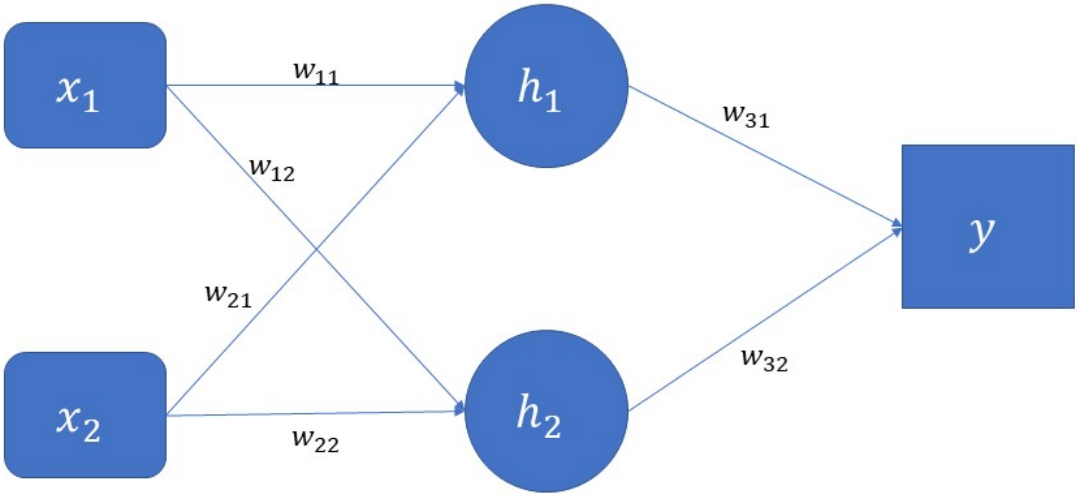
Case4: 1*1+1*1=2

So, for the Case4 the activation function of perceptron $y = \varphi(h)$ will be activated as its value is 2 which is equal to c.

c) (20 points). The logical XAND function (i.e., the Exclusive AND) returns TRUE only when both arguments are the same (e.g., both TRUE or both FALSE). Otherwise, it returns FALSE. This can be represented by the following table.

x_1	0	0	1	1
x_2	0	1	0	1
x_1 XAND x_2	1	0	0	1

Consider a XAND neural network which has two neurons in a single hidden layer.



In the above diagram, $h_i = (w_{i1}x_1 + w_{i2}x_2) \geq c_i$, $i = 1{,}2$ and $y = (w_{31}h_1 + w_{32}h_2) \geq c_3$. Specify the six synaptic weights and the three threshold values such that the above neural network can implement the XAND function. The parameters are still integers but we allow negative integers. You have to prove that your solution does work.

The Possible values for the parameter will be:

w11=1, w21=1, c1=2

w12=-1, w22=-1, c2=0,

w31=1, w32=1, c3=1

Proof:

For: $w_{11}x_1 + w_{21}x_2$

Case:1 1*1+1*1=2

Case:2 1*1+1*0=1

Case:3 1*0+1*1=1

Case:4 1*0+1*0=0

C1 will be activated for Case1 as its value is equal to c1.

For: $w_{12}x_1 + w_{22}x_2$

Case:1 -1*1-1*1=-2

Case:2 -1*1-1*0=-1

Case:3 -(1*0) -(1*1) =-1

Case:4 -(1*0) -(1*0) =0

C2 will be activated for Case4 as its value is equal to c2=0.

For: $w_{31}x_1 + w_{32}x_2$

The C3 will be activated only if when c1 and c2 is activated simultaneously for one case. So:

Case: 1*2+1*0=2