CS 584-04: Machine Learning

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.

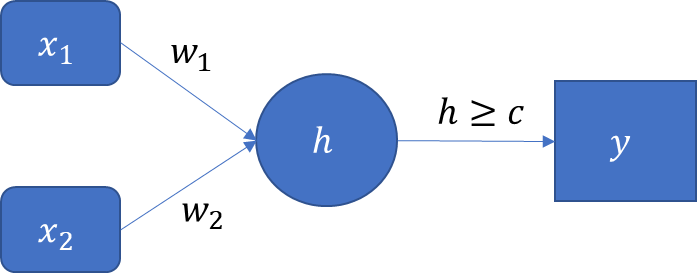
1. (5 points). What is the maximum number of principal components that you can get?
2. (5 points). Plot the Explained Variances against their indices. Add a horizontal reference line whose value is the reciprocal of the number of variables. Label the axes and add grid lines to the axes.
3. (5 points). 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?
4. (5 points). What is the cumulative explained variance ratio accounted by the major principal components that you selected in c)?
5. (5 points). Plot the Elbow and the Silhouette charts against the number of clusters.
6. (5 points). What is the number of clusters that you will choose based on the charts in e)?
7. (5 points). List the names of the communities in each cluster.
8. (5 points). What are Chicago’s annual crude hospitalization rates from 2000 to 2011? Please present your answer in a table.
9. (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.
10. (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?

# 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:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 0 | 1 | 1 |
|  | 0 | 1 | 0 | 1 |
| OR | 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: if . Otherwise, .

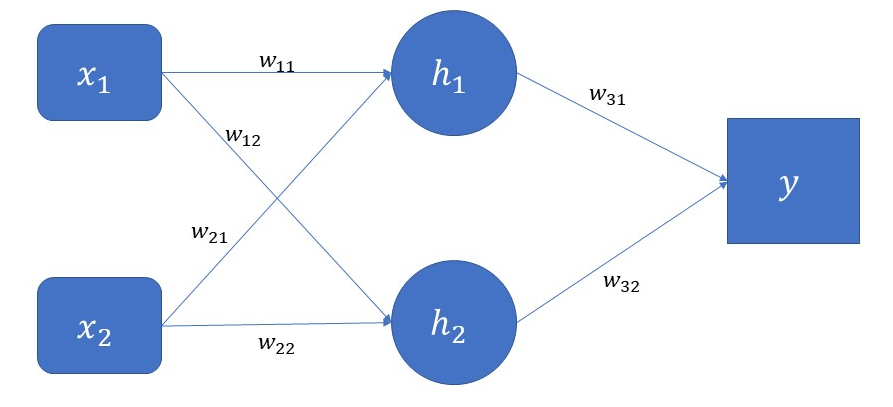
1. (15 points). If we restrict the values of the parameters , , and to positive integers, then specify the lowest possible values for these parameters such that the perceptron can implement the logical OR function. You have to prove that your solution does work.
2. (15 points). If we restrict the values of the parameters , , and 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 0 | 1 | 1 |
|  | 0 | 1 | 0 | 1 |
| AND | 0 | 0 | 0 | 1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 0 | 1 | 1 |
|  | 0 | 1 | 0 | 1 |
| XAND | 1 | 0 | 0 | 1 |

1. (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.

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



In the above diagram, , and . 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.