CSC 539 Computational Epidemics
Summer 2022
Jackson State University
Exam 5
Instructor: Dr. Natarajan Meghanathan

Total Points: 140 **Due: July 28th, 2022, 11.59 PM in Canvas**

Q1 - 15 pts) Covid-19 Testing: Impact of Lab Accuracy. A lab has been testing patients for Covid with one of the two possible results: positive or negative. The lab guarantees that their results are 98.5% accurate :

- i.e., if you have the disease, the result will be positive in 985 of 1000 tests done;
- likewise, if you do not have the disease, the result will be negative in 985 of the 1000 tests done.

Let 5% of the population actually have Covid.

If the test taken for a person gives a positive result, what is the probability that the person has Covid?

Q2 - 20 pts) Predicting Covid-19 using the Symptoms. Consider the following dataset of the symptoms and whether or not the persons were actually diagnosed with Covid-19.

X1	X2	X3	X4	
Sore throat	Cough	Runny nose	Fever	Covid-19?
Yes	Yes	No	Yes	Yes
Yes	No	Yes	Yes	Yes
No	No	Yes	Yes	No
No	Yes	No	Yes	No
Yes	Yes	Yes	No	Yes
Yes	No	No	No	No
Yes	Yes	Yes	Yes	Yes
Yes	No	No	Yes	Yes

Use the Naive Bayes Classifier to predict whether a person who has the following symptoms could have Covid or not?

Sore throat = Yes
Cough = Yes
Runny nose = No
Fever = No

Q3 - 25 pts) Delta vs. Omicron vs. Flu. We have data on 2000 patients who were diagnosed being infected with one of the following: the Delta variant of Covid-19, Omicron variant of Covid-19 or the Influenza virus (flu).

Type	Shortness of Breath		Loss of Taste/Smell		Sore throat		Hoarse Voice		Runny Nose		Fever		Total
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	
Delta: Covid-19	550	150	500	200	400	300	50	650	500	200	500	200	700
Omicron: Covid-19	100	800	50	850	850	50	800	100	600	300	700	200	900
Influenza: Flu	25	375	0	400	350	50	0	400	350	50	350	50	400
Total	675	1325	550	1450	1600	400	850	1150	1450	550	1550	450	2000

Consider a person with the presence or absence of the symptoms as follows:
Shortness of Breath = No
Loss of Taste/Smell = No
Sore Throat = Yes
Hoarse Voice = Yes
Runny Nose = No
Fever = Yes

Use the Naive Bayes Classifier to predict the most likely virus infection the person is to have.

Q4 - 20 pts) CRP Levels: Bacteria vs. Virus Infections. C-Reactive Protein (CRP) is a protein whose concentration in the blood increases when there is inflammation in the body in response to infections due to bacteria, virus, etc. The body temperature is also high in case of infections.

The CRP levels and raise in body temperature are more in case of bacterial infections (class 1) compared to virus infections (class 0).

The following dataset shows the CRP levels and body temperature (measured in F) of people infected with bacteria (class 1) and virus (class 0).

	CRP (X1)	Temp, F (X2)	Actual Class
1	40	96.8	0
2	11.1	98.96	0
3	30	97.7	0
4	21.4	102.92	0
5	10.7	103.28	0
6	3.4	105.26	0
7	42	99.68	1
8	31.1	107.96	1
9	50	101.3	1
10	60.4	102.92	1
11	45.7	101.48	1
12	17.3	108.86	1

(a) Run the logistic regression algorithm on the given training dataset and determine the equation (parameter values) for a linear decision boundary that separates the two classes.
(b) Compute the ΘX values and the hypothesis $h(\Theta X)$ values for logistic regression on the training dataset and show the predicted class for each record in the training dataset.
(c) Determine the coordinates where the equation representing the linear decision boundary of logistic regression cuts the CRP (X1)-axis and the temperature (X2)-axis. Using these coordinates, draw the linear decision boundary over the distribution of the data points in the training dataset and show the separation of the data points into the two classes representing the virus and bacterial infections.

Q5 - 25 pts) Biomarkers for predicting the Pathogenesis of Covid-19. Neutrophils are a critical type of white blood cells that are activated to fight against infections. Chemokines are signaling proteins secreted by cells to activate the neutrophils. During the Covid-19 pandemic, the concentrations (measured in pg/ml) of chemokines such as IL-8 and TNF-α in blood were observed to be effective biomarkers to predict in advance the pathogenesis of the disease (categorized as severe-typically leading to death or at least an ICU admission; moderate - require hospitalization, but no death and mild - does not require hospitalization). The following dataset was extracted from one such clinical study on these biomarkers for Covid-19.

IL-8	TNF-α	Covid-19
25	40	Mild
65	49	Moderate
40	55	Mild
75	55	Moderate
100	80	Severe
125	88	Severe
30	45	Mild
70	55	Moderate
90	70	Severe
55	50	Moderate

Employ logistic regression to fit a multi-class classification model (use the one vs. rest approach of binary classification) and determine the model parameters for predicting each of the three classes.

Use the model to predict the pathogenesis (mild, moderate or severe) of Covid-19 for the following test cases:

IL-8: 46 pg/ml and TNF-α = 44 pg/ml
IL-8: 82 pg/ml and TNF-α = 68 pg/ml

Q6 - 35 pts) Contact Tracing. The DBSCAN clustering algorithm can be used for contact tracing. The following dataset comprises of the locations of five different people (A, B, C, D and E) during a one-week period. If A is found to test positive for Covid-19, find who else among B, C, D and E were in the vicinity of A: i.e., in the same DBSCAN cluster(s) as A, during the one-week period and could be considered as vulnerable contacts?

Person	Data points	X1	X2
A	1	35	27
D	2	10	27
C	3	25	27
D	4	18	29
E	5	9	29
D	6	3	30
A	7	35	28
C	8	26	31
B	9	42	28
C	10	50	29
B	11	38	29
E	12	14	32

a) Show the normalized values for X1 and X2.
b) Show the pair-wise distance matrix based on the normalized scale.
c) Choose an appropriate value for the parameter minPts (k) and show the k-avg distances for the data points.
d) Show a plot of the sorted k-avg distances of the data points and identification of parameter eps(ε) using the elbow method.
e) Show the pair wise distance matrix wherein entries with values less than eps(ε) are retained as the ε-neighbors of the data points.
f) Using the value of $1 + |\varepsilon\text{-neighborhood}|$ size, classify the 12 data points as core (if $1 + |\varepsilon\text{-neighborhood}| \geq \text{minPts}$) and non-core (if $1 + |\varepsilon\text{-neighborhood}| < \text{minPts}$).
g) Using the core data points and the ε-neighborhood of all the data points as the basis, determine one or more clusters of the data points as well as identify the outliers, if any.
h) Using the results of (g), identify people who were with A and need to be quarantined.