Miniproject 3
BMI 555 IEE 520
Fall 2020

Due Date: October 22, 2020

1) A support vector machine is trained to classify a data set with N = 10,000 instances. The following data is a subset of (standardized) selected rows from the training data set.
Support vectors are indicated with the column alpha and *no additional* support vectors occur.
The kernel was a Gaussian kernel with gamma = 2.0.
Complete the following calculations without a support vector machine package, so you understand the steps. You might use *Microsoft Excel*.

Instance	x1	x2	x3	x4	alpha	y	
1	0.15	0.50	-0.66	-0.18	0	1	
2	-1.24	-0.75	-1.45	-1.04	0	-1	
3	0.48	0.47	-0.10	0.27	0.1	1	
4	0.51	0.71	-0.41	0.13	0.2	-1	
5	-0.63	-0.79	-1.27	-0.22	0	1	

a) Compute f(x) for each instance of data.

b) To what class is each instance assigned by the model?

c) If C is reduced from 10 to 1, do you expect more or fewer errors on the training data? Why?

2) A support vector machine is trained to classify a data set with N = 10,000 instances and 3 classes, denoted as {A, B, C}. The following is a summary of the assignment of a collection of *one versus the rest* classifiers for several instances of data. What is the assigned class for each instance? In case of a tie, list all possible classes.

Instance	A versus rest	B versus rest	C versus rest	Assigned class(es)?	
1	A	rest	rest		
2	rest	rest	C		
3	A	rest	C		
4	A	B	rest		

3) Consider a nearest neighbor classifier with Euclidean distance for the training data in the first exercise. Complete the following calculations without a nearest neighbor package, so you understand the steps. You might use *Microsoft Excel*.

a) If $K = 3$, to what class is the following instance assigned? Explain.

$x1$	$x2$	$x3$	$x4$
0.06	0.51	-0.31	-0.24

b) Suppose a Gaussian kernel function exp(-0.5$||xi - x0||^2$/lambda) with lambda = 2 is used for weights. Show your work used to assign the instance.

c) If K is increased from 3 to 5, do you increase the decrease the complexity of the model?

4) Use the *diabetes* dataset. Use the SVM package in SKLEARN ONLY, to design the best network for me to use on similar data in the future.
But I am interested in a balanced error rate between the two classes, so set weight = 'balanced' in SKLEARN.
Also, I want to know about the quality of your model. Provide an estimate the generalization error of the model you recommend.
Submit

- Your final code, your output, and your estimate of generalization error
- A brief description of what models you explored and how you estimated generalization error.
- A clear statement of parameters changed from the SKLEARN defaults. It is important to learn to communicate your models clearly.

You are writing for me, so you can summarize based on my knowledge of the topic.