CSCI4380/6380 - HW1

Instructor: Dr. Ninghao Liu (ninghao.liu@uga.edu)

January 19, 2022

– Upload two files to eLC:
1) a scanned handwritten solution or typed pdf file named "*YourID_HW1.pdf*" containing your answer to each question;
2) a zip file named "*YourID_HW1PQ.zip*" containing your programs for Question 2.

– Due Date: February 8, 2022

1 kNN Classifiers (20pt)

Given the following dataset:

No.	x_1	x_2	Class Label
1	-1	1	0
2	-1	2	0
3	-1	4	0
4	-2	3	0
5	1	1	1
6	1	3	1
7	2	2	1
8	3	2	1
9	0.6	3.5	?

1) (8 pts) Classify instance No. 9 using the kNN classifier with $k = 3$. Use squared distance (i.e., Euclidean distance, L2 distance) as the distance measure between two instances.

2) (6 pts) Draw the decision boundary of kNN classifier with $k = 1$. Only consider the instance NO. 1, 2, 5 in the table above as training data.

3) (6 pts) Explain at least two disadvantages of kNN classifiers.

2 Linear Models (20pt)

1) (10 pts) Given a dataset with three one-dimensional instances $\{([2], 4), ([4], 2), ([3], 2.5)\}$ (e.g., the feature vector and ground-truth label of the first instance is [2] and 4, respectively). Suppose we use a linear predictor $f = w \cdot x + w_0$ to fit the data (w_0 is called the $offset$), with squared loss as the loss function. Complete the codes in "hw1_gradientDescent.py" to solve w and w_0 using gradient descent.

Report in "*YourID_HW1.pdf*": (1) the codes you write to fulfill the functionality; (2) the outputs at iteration 0**,** 1990 **and** 2000**.**

2) (10 pts) With the same problem settings above, complete the codes in "hw1_stochasticGradientDescent.py" to solve w and w_0 using stochastic gradient descent.

Report in "*YourID_HW1.pdf*": (1) the codes you write to fulfill the functionality; (2) the outputs at iteration 0**,** 5900 **and** 6000**.**

3 Naive Bayes Classifiers (30 pts)

Given the following dataset (each instance has three features):

No.	Outlook	Temperature	Humidity	Play Golf
1	sunny	hot	high	N
2	sunny	hot	high	N
3	overcast	hot	high	Y
4	rain	mild	high	Y
5	rain	cool	normal	N
6	rain	cool	normal	N
7	overcast	cool	normal	Y
8	sunny	cool	normal	Y
9	sunny	mild	high	?

1) (15 pts) Classify instance No. 9 using Naive Bayes Classifier (NBC). Include the details of your NBC, probability calculations, and how the final classification is determined.

2) (5 pts) What is the time complexity for training and testing Naive Bayes classifier, respectively?

3) (10 pts) After a yearly checkup for a software developer, there are both bad news and good news from the doctor. The bad news is that the developer has a test result positive for a serious disease, and the test is 98% accurate (i.e., if you have the disease, then the probability of testing positive is 0.98; if you do not have the disease, the probability of testing negative is also 0.98). The good news is that this is a rare disease, because only 1 in 20,000 people will have it. What are the chances that the developer actually has the disease?

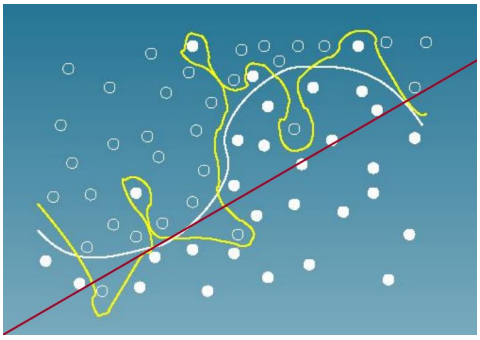


Figure 1: Overfitting and Underfitting.

4 Classification - General (15 pts)

1) (5 pts) Introduce the main idea of k-fold cross validation. (Using illustrative figures could be helpful.)

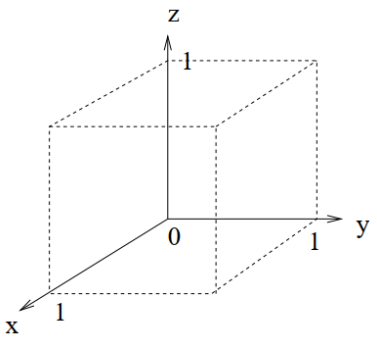
2) (5 pts) Given the three classification boundaries in Figure 1 specified by the red, yellow and white curves respectively, explain which line is most likely to correspond to the overfitting classifier? Which is most likely to correspond to the underfitting classifier?

3) (5 pts) For kNN classifiers, explain the relationship between parameter k and the model's tendency to overfitting.

5 Neural Networks: MLP (10 pts)

Can a single perceptron unit solve the classification problem in the below table?: In other words, can we find a set of weights for a single perceptron unit to correctly classify all examples? (1) Answer "yes" or "no" to the question (draw a 3D geometric illustration of the problem may help you understand the problem); (2) justify your reasoning.

Input x	Input y	Input z	Class
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0



6 Distance Metric (5 pts)

Prove that Euclidean distance is a valid distance metric.
(Hint: Cauchy's Inequality. $\sum_{i=1}^{N} r_i^2 \sum_{i=1}^{N} s_i^2 \geq (\sum_{i=1}^{N} r_i \cdot s_i)^2$ for all $r_i, s_i \in \mathbb{R}$.)