CS410: Artificial Intelligence 2020 Fall
Homework 2: Machine Learning Basics and Logistic Regression
Due date: 23:59:59 (GMT +08:00), November 11, 2020

1 True or False Questions

1. Classification and regression are both supervised learning problems.

2. Overfitting may occur in both classification and regression problems.

3. Generally, regression is not used in classification problems, but there are also special ones. For example, logistic regression can be used to solve 0/1 classification problems.

4. The most commonly used evaluations for both regression and classification problems are precision and recall.

5.(single-choice) After evaluating the model, it is concluded that there is a bias in the model. Which of the following methods may solve this problem:
A. Reduce the number of features in the model
B. Add more features to the model
C. Add more data
D. Both B and C
E. All of the above

6.(single-choice) To do a two-class prediction problem, first set the threshold to 0.5. Samples with a probability greater than or equal to 0.5 are classified as positive examples (i.e. 1), and samples less than 0.5 are classified as negative examples (i.e. 0). Then, use the threshold $n(n > 0.5)$ to re-divide the sample into positive and negative examples. Which of the following statements is correct ()
(a) Increasing the threshold will not increase the recall
(b) Increasing the threshold will increase the recall
(c) Increasing the threshold will not reduce the precision
(d) Increasing the threshold will reduce the precision

A. (a) B. (b) C.(a) and (c) D.(b) and (d) E. None

2 Calculation Questions (Please provide the detailed calculation process)

1.Given 100 human photographs with 50 female (labeled as 1) and 50 male (labeled as 0), an algorithm predicts 45 to be male, 55 to be female. Among the 45 male predictions, 5 predictions are not correct. Among the 55 female predictions, 10 predictions are not correct.
Please calculate the Accuracy, Precision, Recall and F-measure of this algorithm.

2. Assuming that the probability of a series of samples being divided into positive classes has been obtained, the following figure is an example. There are 10 test samples in the figure. The "Class" column indicates the true label of each test sample (1 represents positive sample, 0 represents negative sample), "Prediction" represents the probability of each test sample belonging to a positive sample.
Please calculate the AUC score.

Class	Prediction
1	0.92
1	0.88
0	0.76
1	0.67
1	0.59
1	0.43
0	0.38
0	0.36
1	0.29
0	0.2

3 Short Answer Questions

1. How to judge whether overfitting has occurred?
2. What are the reasons for overfitting and how to solve it?
3. Please list common regularization methods and their comparison.
4. Please describe the difference between generative and discriminative models.

4 Programming

1. National Institute of Diabetes and Digestive and Kidney Diseases wants to establish a model that automatically diagnose whether a Pima Indian has diabetes and provides a dataset as the material of the establishment of the model (Please see Supplementary materials). The dataset contains information like Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, of female Pima Indians over 21. In the dataset, outcome labels as 1 means that the person has diabetes. Otherwise, she does not have.

Please implement Logistic Regression to establish this model, and use gradient descent to optimize the model. Change the learning rate gradient descent to plot accuracy-learning rate graph.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1