Machine Learning HW 8

1. Training accuracy: 0.7742857142857142. Testing accuracy: 0.7257142857142858
2. The features of our data we are using are generated by conditionally independent distribution, meaning that a presence of a specific feature in a class is not related to the presence of any other feature. In the context of credit scores, all the features contribute independently to the probability that a data is classified as “good” or “bad” credit.
3. For the “month” attribute, we need to make a different interval of several month attributes (kind of like creating a one-hot vector) so that we can represent a single “month” attribute as several attributes that represents the interval in which the original “month” attribute is in. For the “Credit Amount” attribute, we also create several different attributes that classify the original “credit amount” values based on the interval. For the “Number of credit” attributes, we make a one hot vector with size 4 because the feature value has four variations. For the “Credit” attribute, we just have to change it as 2 = 1 and 1 = 0.
4. Disparate Impact is one of the criterions to measure the fairness of a machine learning algorithm. Specifically, the disparate impact considers the prediction model to be fair if the following equation holds. (In this example, we see t = 0.8 as it is a common threshold used in the U.S.) In the inequality shown below, Y’ represents the hypothesis’s output and S represents whether the person is in a protected group. If the practice does not result in the following formula, it would be considered as illegal discrimination suggesting it has a disproportionately adverse effect on members of a protected group. Overall, this method uses a useful measure because it statistically reveals a direct relationship between a particular feature of a group and a hypothesis’s output. However, this method is not always necessarily great because even if there is actually a fair relationship between a particular feature and an output label, the metric could still suggest that the data is unfair, which could ironically result in neglecting strengths of a particular group.

$$P[\hat{Y} = 1|S! = 1] \ / \ P[\hat{Y} = 1|S = 1] <= (t = 0.8)$$

5. In our context, false positive rate means the probability of our model predicts a person to have a good credit when a person in fact has a bad credit. On the other hand, a false negative rate means the probability of our model predicts a person to have a bad credit when a person in fact has a good credit. If one group’s FPR is much higher than other groups, it means that particular group of people with bad credit are more likely to be predicted to have a good credit even if they don’t, suggesting that the feature that the group is sharing is probably an unfair feature to be used for predicting a credit score. This could ultimately result in giving a privilege to particular people and a serious problem accordingly. On the other hand, if a particular group’s FNR is much higher than other groups, it means that that group of people with good credit are less likely to be predicted to have a good credit, also suggesting that the feature they have in common is an unfair feature to be used for the prediction of credits. This could lead to yet another problem as a particular group of people could be looked down on or have disadvantage even if they have good credit or qualified to have one.