CSE 474: Introduction to Machine Learning

Programming Assignment - 3

Naive Bayes Classifier

Group 17:

Rupali Kotina

Gitanjali Nandi

Rajeev Gundavarapu

Introduction

In this project, we have implemented a Naive Bayes Classifier and tested it on the german.data.pickle file, which contains mock credit application data of customers. The Naive Bayes Classifier that was implemented utilizes priors on the target class variable as well as individual features.

Variation of the Cross-Validation FScore

We incorporate the use of cross-validation to measure the performance of the Naive Bayes Classifier on the provided data set by splitting the given data into both training and test data. We make use of the F-measure which is the harmonic mean of number of correctly classified test instanced of the class being considered divided by actual number ot test instances of the class.

The FScore for the given parameters a = 1, b = 1, alpha = 1 is 0.69.

The following table shows the FScore that was obtained by varying a,b, and alpha values from 0.5 to 5 in intervals of 0.5:

Value of Variation	FScore(Varying a)	FScore(Varying b)	FScore(varying alpha)
0.5	0.69	0.69	0.69
1	0.69	0.69	0.69
1.5	0.69	0.69	0.70
2	0.69	0.69	0.70
2.5	0.69	0.69	0.69
3.0	0.69	0.69	0.69
3.5	0.69	0.69	0.69
4.0	0.69	0.69	0.69
4.5	0.69	0.69	0.69
5.0	0.69	0.69	0.69

It can be observed that there is no significant difference in variation for either of the parameters, but alpha slightly increases from values 1.5-2. This shows that the prior probabilities are not having much of an effect on the FScore value. However, it can be noted that although the FScore Value does not change significantly even for larger numbers such as 25, alpha shows a considerable difference with a decreasing Fscore.

Value of Variation	FScore(Alpha)
20	0.61
25	0.60
30	0.58
35	0.57
40	0.55
45	0.55
50	0.53
55	0.52

Disparate Impact and Sensitive Features (Age and Gender)

We measure the fairness of the Naive Bayes Classifier algorithm with respect to the Disparate Impact Value, for which a higher values indicates that the algorithm is unfair to the unprivileged customers. Here we are considering the sensitive features, age and gender. We are checking if the algorithm is unfair to the younger customers when we are considering age as the sensitive feature. If we are considering gender, we are measuring if the algorithm is unfair to the female customers. In this case, a Disparate impact greater than 1 implies that the classifier is unfair against the unprivileged customers.

Sensitive Feature	Disparate Impact
Age(si=0)	1.10
Gender(si=1)	1.25

Based on the DI values, it can be noted that the Naive Bayes Classifier Algorithm is unfair towards the younger customers and the female customers considering the sensitive features age and gender respectively.

Disparate Impact of Classifier Based on Bias

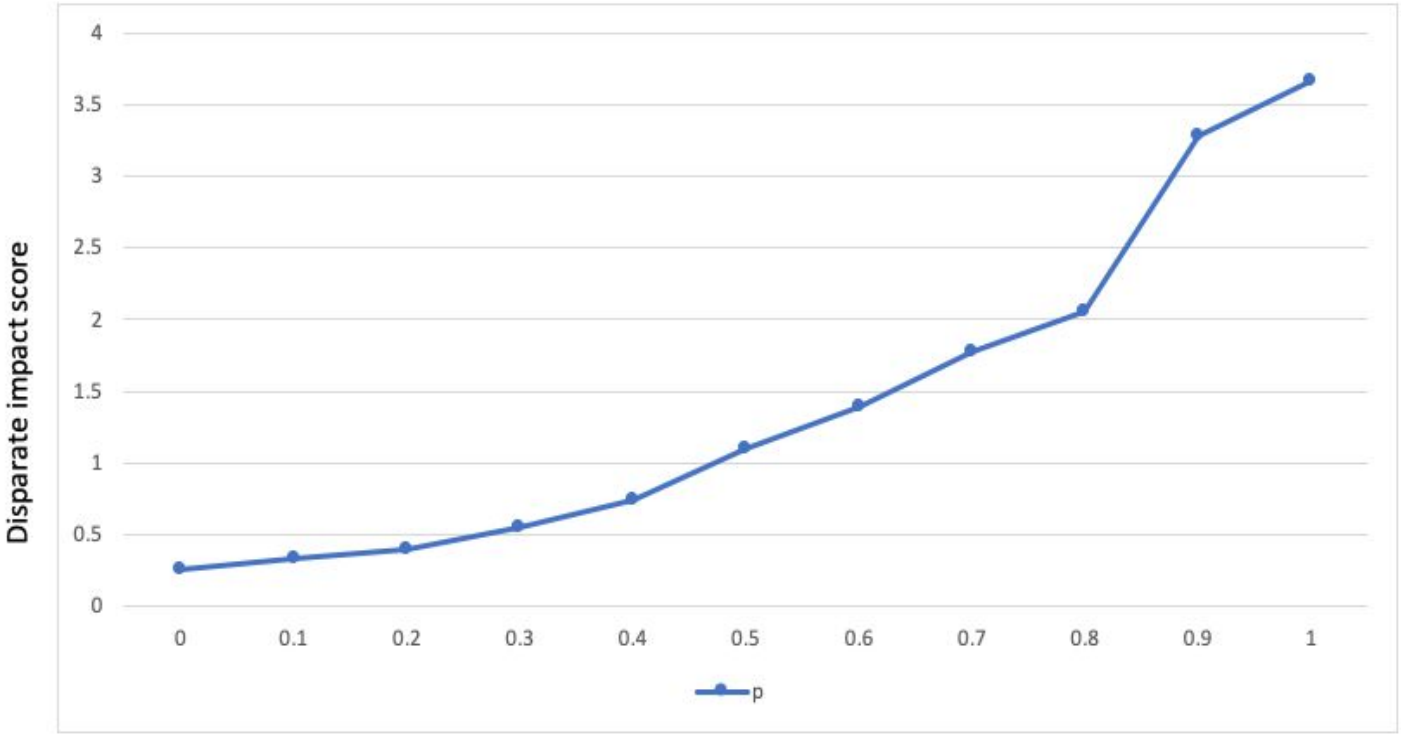
The classifier sometimes tends to be unfair due to the bias present within the training data itself. We make use of the function genBiasedSample which induces an artificial bias in the data set by oversampling bad examples from underprivileged set of customers from either of the sensitive features, age (si = 0) and gender (si = 1). In the function p represents the probability of picking a bad customer from the unprivileged set. Below, we measure the disparate impact of the classifier as a function of p for the sensitive features age and gender.

For the Sensitive Feature Age:
si = 0
This represents the population which is under 25 years of age as underprivileged. The following table below shows the values of the Disparate Impact by varying the value of p passed into the genBiasedSample function which induces an artificial bias.

p(bias)	Disparate Impact Score
0	0.25
0.1	0.33
0.2	0.40
0.3	0.55
0.4	0.74

0.5(no bias)	1.10
0.6	1.39
0.7	1.78
0.8	2.06
0.9	3.28
1.0	3.66

The distribution of p vs DI Score is shown below:

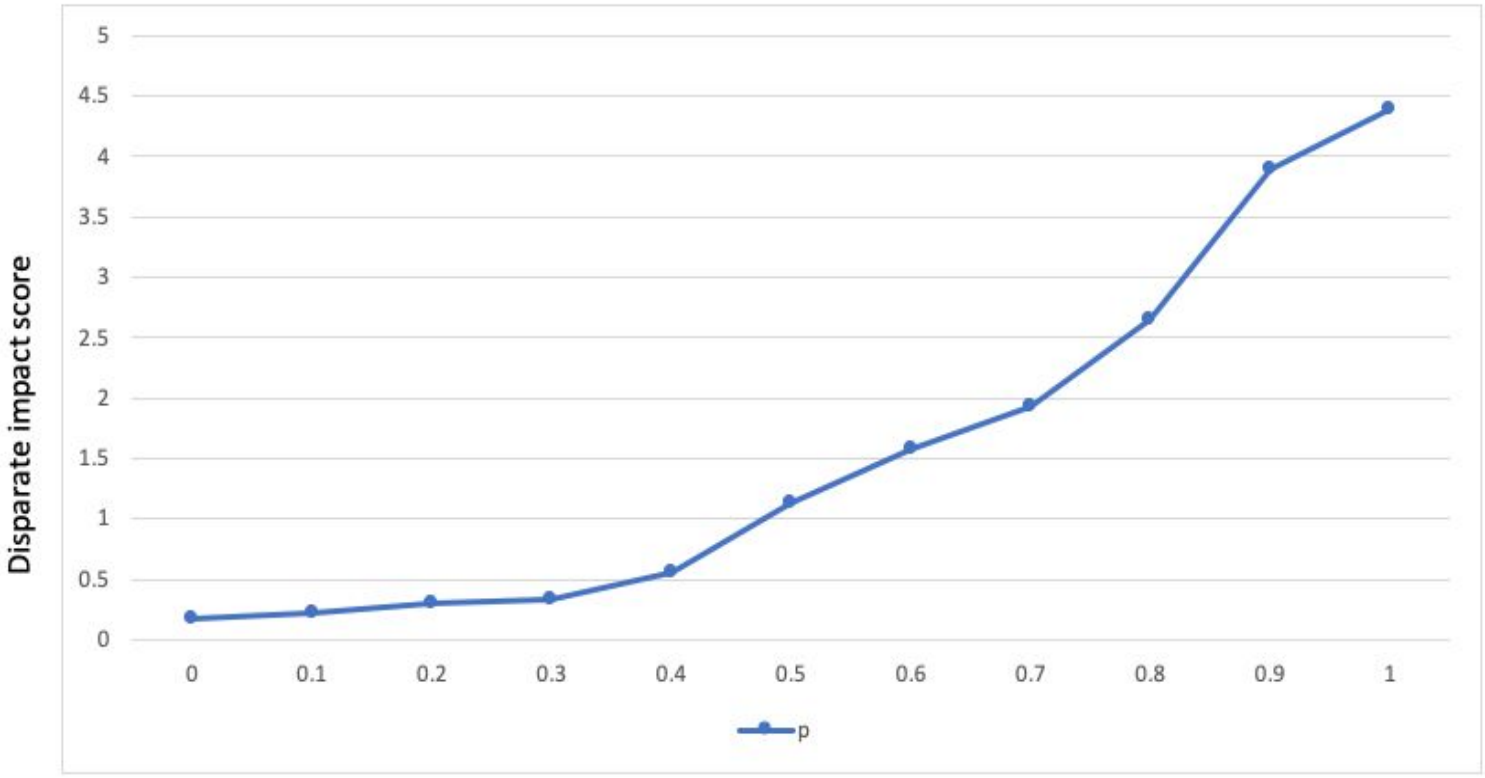


In the following graph, it can be observed that for $p < 0.5$ the DI value is less than 1. This shows that the Disparate Impact is biased against the privileged customers, which are in this case, those customers who are of age greater than or equal to 25 years. When $p = 0.5$ it can be noticed that the Disparate Impact value is 1, which shows that the classifier is unfair as it is biased against the underprivileged customers. When $p>0.5$ the graph is more biased against the underprivileged customers($DI > 1$).

For the Sensitive Feature Gender:

p(bias)	Disparate Impact
0	0.17
0.1	0.22
0.2	0.31
0.3	0.34
0.4	0.56
0.5	1.13
0.6	1.58
0.7	1.93
0.8	2.65
0.9	3.89
1.0	4.39

The distribution of p vs DI score for the above chart is shown below:



From the graph, it can be seen that when $p < 0.5$, the the disparate impact score is biased against the privileged customers, which in this case happens to be the male customer population based on the DI value which is less than 1. The graph at $p = 0.5$ is biased against the female customers which are underprivileged ($DI > 1$). When $p > 0.5$ there exists a bias which is heavily against the underprivileged female customer population as the bias is increased towards 1($DI > 1$).