Rohan Salvi

Homework 4



(Number of correct predictions/number of predictions made) x 100 =

= (12/20) x 100

= 60 %

1. The result of a very basic model/solution is a baseline. To get a better result, we usually start with a simple answer and work our way up to more sophisticated solutions. It's good if we get a higher score than the baseline.
2. We can't predict which algorithm will perform best for our problem ahead of time, so we'll have to try a few of them and see which one works best, then focus on that one. As a result, when working on a machine learning challenge, it's vital to establish a baseline of performance. A baseline serves as a benchmark against which other machine learning algorithms can be measured.

d)

To compare the performance of our SVM classifier to, we can utilize Naive Bayes. It is simple and quick to forecast the test data set's class. It's also good at multi-class prediction. When the assumption of independence is true, a Naive Bayes classifier outperforms alternative models such as logistic regression, and it requires less training data.

e)

Accuracy is the state or attribute of being true, right, or exact, it is the absence of mistake or fault. Although the proportion of right classifications among all classifications is a straightforward and "intuitive" statistic, it may be a bad measure for data that is unbalanced.

2.

a)

No. of True +Ve = 2

No. of False -Ve = 5

No. of True -Ve = 10

b) Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | 2 | 3 |
| Negative | 5 | 10 |

c)

p(Y,p)= 2/20 =0.1

p(Y,n) =3/20 = .15

p(N,p) =5/20 = 0.25

p(N,n) =10/20 =0.5

d)

*expected benefit =p*(*Y,p*)*×b*(*Y,p*)+*p*(*N,p*)*×b*(*N,p*)+*p*(*N,n*)*×b*(*N,n*)+*p*(*Y,n*)*×b*(*Y,n*)

(0.1x10000 )+0.25 x (-15000) +0.5 x 5000 +0.15 x (-3000) = -700 $

3.

a. The profit curve is a graph that shows how much profit a company can make by producing a certain amount of production. A profit curve may be simply calculated from a table of profit and output quantity figures.

b.

|  |  |
| --- | --- |
| ROC | Profit Curve |
| ROC curves are widely used to depict the relationship/trade-off between clinical sensitivity and specificity for each conceivable cut-off for a test or a set of tests in a graphical format. | The profit curve is a graph that shows how much profit a company can make by producing a certain amount of output. |
| Model correctness is represented by the area under the curve. The greater the surface area, the greater the precision. If the curve on the graph spreads to its edges, the model is 100 percent accurate. | The greatest profits for and target population are represented by the curve's peaks. |

c.

The first major distinction is that accuracy is calculated on anticipated classes, whereas ROC is calculated on predicted scores. That implies we'll need to figure out what the best threshold is for our problem. Furthermore, accuracy takes into account the percentage of correctly assigned positive and negative classes.

d.

If the Roc value is 0.5 then , the classifier can't tell the difference between the two +ve and -ve classes. For all of the data points, the classifier is either predicting a random class or a constant class.

e.

We will utilize ROC curve and cross validation metrics because we will be utilizing SVM classifier. To create a ROC curve from your data, first rank all the numbers and then correlate each value to a diagnosis. We'll make it by computing and showing the true positive rate vs the false positive rate for a single classifier at various levels. We can use it define breast cancer data set.