CS 613 - Machine Learning

Assignment 2 - Classification - Abishek S Kumar

Solutions

This is the solutions pdf file of the Assignment 2

Answers to Question 1

1. Based on the total counts in the given table for Counts:

   (a) The total entropy for the dataset:

   $$H(Y) = -\sum_{i=1}^{n} P(v_i) \log_2(v_i)$$

   i. Total no. of samples, Summation of counts = 21

   ii. Total no. Positive Counts= 12

   iii. Total no. Negative Counts = 9

   The total entropy is:

   $$H(Y) = -(12/21) * \log_2(12/21) - (9/21) * \log_2(9/21)$$

   $$H(Y) = 0.98$$

   (b) The Information Gains and seperate feature entropies are calculated as follows:

   $$E(H(A))_{x_1} = \sum_{i=1}^{n} (p_i + n_i/p + n) H(P(v_i)....P(v_n))$$

   Entropy for feature x_1 :

   $$E(H(A))_{x_1} = \frac{8}{12}(-\frac{7}{8}\log_2\frac{7}{8} - \frac{1}{8}\log_2\frac{1}{8}) + \frac{13}{21}(-\frac{5}{13}\log_2\frac{5}{13} - \frac{8}{13}\log_2\frac{8}{13})$$

   $$E(H(A))_{x_1} = 0.802$$

   The information gain IG is: $IGx_1 = H(Y) - E(H(A))x_1$, $IGx_1 = 0.183$

   Entropy for feature x_2 :

   $$E(H(A))_{x_2} = \frac{10}{21}(-\frac{7}{10}\log_2\frac{7}{10} - \frac{3}{10}\log_2\frac{3}{10}) + \frac{11}{21}(-\frac{5}{11}\log_2\frac{5}{11} - \frac{6}{11}\log_2\frac{6}{11})$$

   $$E(H(A))_{x_1} = 0.94$$

   The information gain IG is: $IGx_1 = H(Y) - E(H(A))x_1$, $IGx_1 = 0.044$

   (c) We choose IGx_1 as the root node as it has the higher information gain, then based on the counts for each possible values of the first feature the tree is as follows:

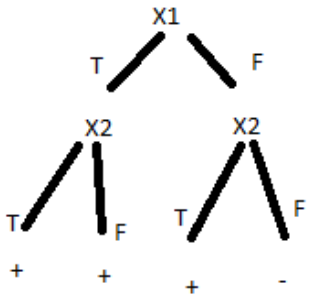

Figure 1: Decision Tree

2. Class Priors calculated:

(a) $P(A = Yes) = \frac{Total no. of samples' Yes'}{Total no. of samples}, P(A) = \frac{3}{5} = 0.6, P(A = No) = \frac{Total no. of samples' No'}{Total no. of samples}, P(A) = \frac{2}{5} = 0.4$

(b) The standardized dataset:

No. of chars	Avg. word length	Decision
0.0615	1.394	Yes
-1.07	0.635	Yes
0.723	-1.447	No
-1.139	-0.730	Yes
1.4243	0.1467	No

(c) Gaussian parameters 'Yes' and 'No' samples:

$$\text{Yes samples: } \mu_{No. of chars} = -0.716, \sigma_{No. of chars} = 0.550$$
$$\mu_{Avg. word length} = 0.433, \sigma_{Avg. word length} = 0.879$$

$$\text{No samples: } \mu_{No. of chars} = 1.074, \sigma_{No. of chars} = -0.650$$
$$\mu_{Avg. word length} = 0.350, \sigma_{Avg. word length} = 0.797$$

(d) Using the gaussian parameters on test sample and the Naive Bayes approach on No. of chars = 242, Avg. word length = 4.56, post standardization, No. of chars = 0.261, Avg. word length =0.450:

$$P(Y = Yes | f_1 = 0.261, f_2 = 0.45) = 0.361$$

$$P(Y = No | f_1 = 0.261, f_2 = 0.45) = 0.108$$

The probability of 'Yes' is higher the classifer will provide a *Yes* for the provided test sample.

3. A validation set can be used to determine what 'k' value fits the training data. Varying 'k' will vary the fitting of the classifier, choosing the lowest value of 'k' would be helpful and thus has to be selected.

Answers to Question 2

1. The final model along with coefficients from gradient descent is as follows: $y = 3.0940x_0 - 4.5744x_1 + 5.6791x_2$ wherein the values of theta are:

   (a) $\theta_0 = 3.0940$

   (b) $\theta_1 = -4.5744$

   (c) $\theta_2 = 5.6791$

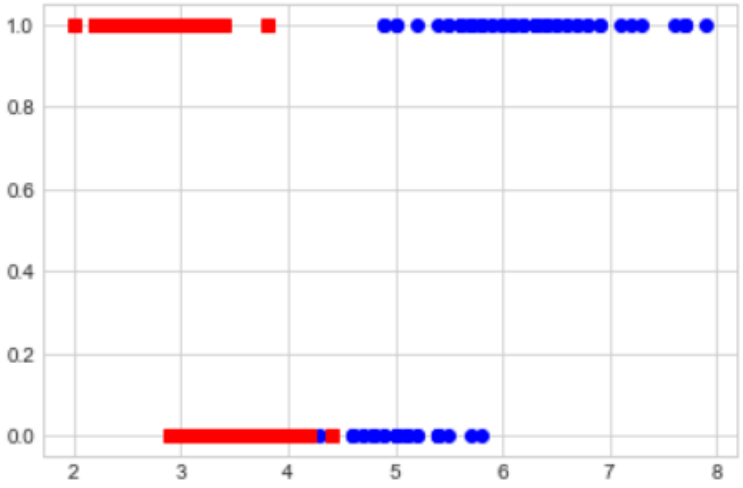

Figure 2: Samples of Features 1 and 2 plot

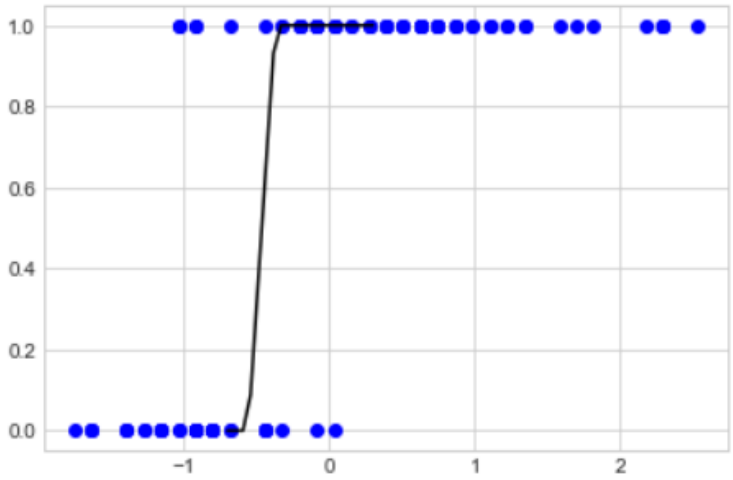

Figure 3: Feature 1 vs. Decision Boundary

2. The final model built from LinearRegression and sklearn library is as follows: $y = 37.7844x_0 - 22.164x_1 + 24.009x_2$ wherein the values of theta are:

   (a) $\theta_0 = 37.7844$

   (b) $\theta_1 = -22.164$

   (c) $\theta_2 = 24.009$

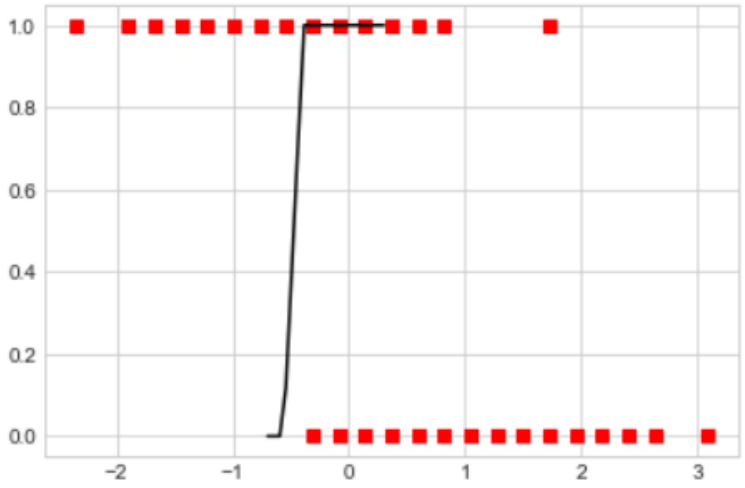

Figure 4: Feature 2 vs. Decision Boundary

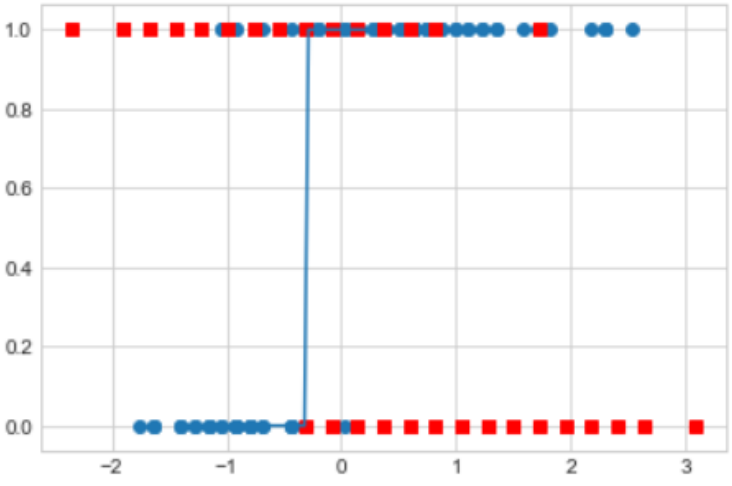

Figure 5: Feature 1 and 2 vs. Sklearn lgr.fit() Decision Boundary

Answers to Question 3

1. $Precision = 0.8765$
2. $Recall = 0.82894$
3. $Accuracy = 0.88479$
4. $F - measure = 0.85207$

Answers to Question 4

1. $Precision = 0.6186$
2. $Recall = 0.9142$
3. $Accuracy = 0.7304$
4. $F - measure = 0.7379$

Answers to Question 5

1. $Precision = 0.9486$
2. $Recall = 0.9639$
3. $Accuracy = 0.9227$
4. $F - measure = 0.9563$