**Week 1 Assignment**

Aneri M. Barbhaya

MPSA, Northeastern University

ALY 6020: Predictive Analytics

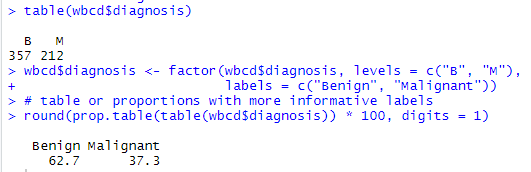
Na Yu

April 11, 2020

**PART-A**

**Working**

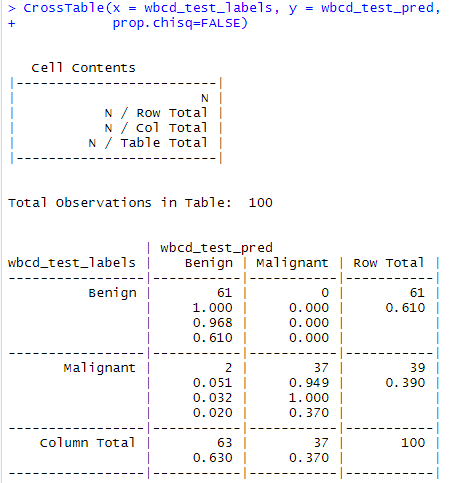
Wisconsin Breast Cancer Diagnostic dataset is being used from the UCI Machine Learning Repository for running KNN algorithm. The csv file is imported in RStudio and data preparation is done using R. The str(wbcd) command shows that the diagnosis column represents factor with 2 levels: “B” and “M”. This is the variable which is to be predicted by learning the measurement values. The other data is represented as numerical features for cancer mass. Since the feature ‘id’ does not provide any significant information about analysis, it is dropped from the dataset. The table function shows that diagnosis column has 357 benign as well as 212 malignant observations. Converting to percentage showed larger benign population. To perform analysis, the diagnosis variable is recoded-converted into factor and appropriately labelled.



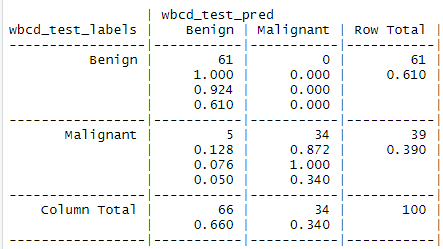
To apply KNN we will first use normalize function. The summary shows the min max values of all the variables in the dataset and this output shows that their ranges differ largely from each other which can cause a problem in classifying or predicting a new data point, so we normalize the numerical features in order to have the same scale for all input variables. Lapply function is used to normalize all the variables. The normalized data is split into train and test datasets which consist of observations 1 to 469 and 470 to 569 respectively; and labels are created for these training and test data. Finally, k-NN is applied with k=21.

Crosstables are used to check the classification according to the labels and evaluate the accuracy of predicted test labels with actual test labels. This includes the verification of the results to see if there were any misclassifications by the model. Top left cell indicates true negative, bottom right cell indicated true positive, bottom left cell indicates false negative and top right cell indicates false positive.

The below cross table shows 61 and 37 values falling under the correct matrix part showing the true positive values, 2 results are false negative (2 tumors actually malignant-predicted benign) and 0 false positive records (actual benign-predicted malignant). 2 out of 100, or 2 percent of masses were incorrectly classified by the k-NN approach i.e. 98% accuracy for the model.

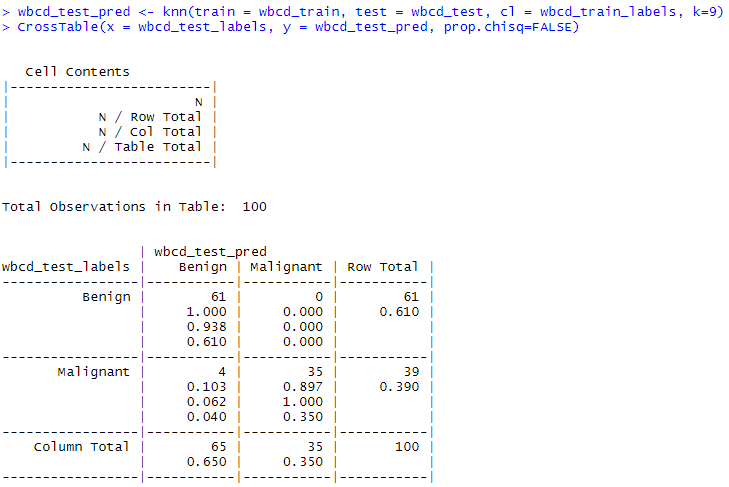


Considering the real-world data nature, certain data might contain higher maximum and very low minimum values. This disturbs the normalization concept and might cause some abnormalities while predicting/classifying as the exteme values are not compressed towards the center. So we might consider using z-score standardization to improve model’s performance because it allows comparison of scores on different kinds of variables by standardizing the distribution. The scale() function is used to scale the features according to the z-score for all the measurement features. To check that the transformation was applied correctly we will use summary (). The mean should always be 0 for z-score standardization. The same process of splitting the data into training and testing sets is performed. KNN is performed on this data and crosstable is applied to check the accuracy and see if the model performance has improved. For k=21, No false positive, 5 false negative. Accuracy 95%. This classifier variation caused a slight decline in accuracy.

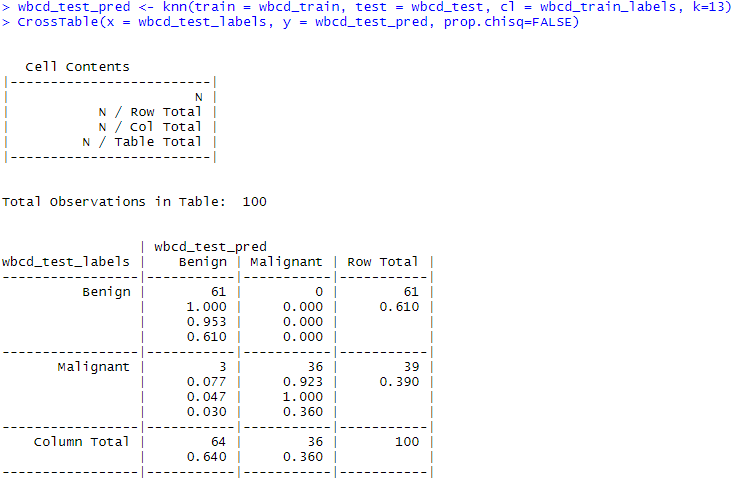


Alternatively, we can test for different values of K to see the variation in model’s performance. For both the models above, we considered k=21. Let’s see for k=9, 13, 25

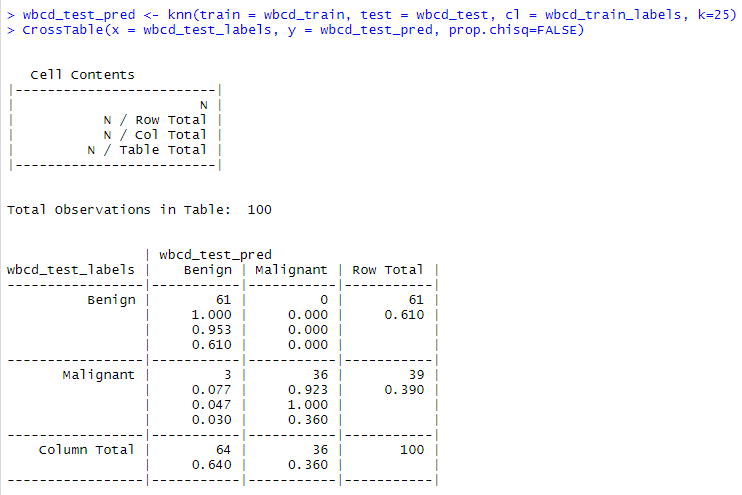
**K=9: 4 false negatives, 0 false positives.**



**K=13: 3 false negatives, 0 false positives**.



**K=25: 3 false negatives, 0 false positives.**



**Testing alternative values of K**

|  |  |  |  |
| --- | --- | --- | --- |
| k value | False negatives | False positives | Percent classified incorrectly |
| 1 | 1 | 3 | 4 percent |
| 5 | 2 | 0 | 2 percent |
| 11 | 3 | 0 | 3 percent |
| 15 | 3 | 0 | 3 percent |
| 21 | 2 | 0 | 2 percent |
| 27 | 4 | 0 | 4 percent |
| 9 | 4 | 0 | 4 percent |
| 13 | 3 | 0 | 3 percent |
| 25 | 3 | 0 | 3 percent |

**Summary/Conclusion**

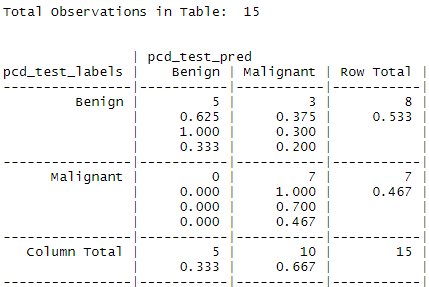
The study of the KNN for the breast cancer example shows the KNN classification approach for the 100 test samples.The results show that there are misclassifications. Classifier was never perfect for any handpicked k value. When k value was 1, false negatives were avoided at the cost of adding false positives. As the k increased, false positives were avoided but false negatives increased at a steady rate. Also tailoring this k-value would be optimal for this sample but would need to be changed for suiting another 100 samples. From the table above, for this test data, a k value of 15-25 would work good.

**PART-B**

For part B, Prostate Cancer dataset from Kaggle is used for KNN classification. The data contains about 100 observations/samples of 10 variables like radius, texture, perimeter, smoothness, area, compactness, symmetry, fractal\_dimension and diagnosis result etc. The same 5 steps of data collection, data exploration and preparation, model training, model performance evaluation and model performance improvement are performed on the data. 90% of data is used in training and 10% constitutes the testing data. Diagnosis\_result is the target variable which is a factor of 2 levels “Benign” and “Malignant”.

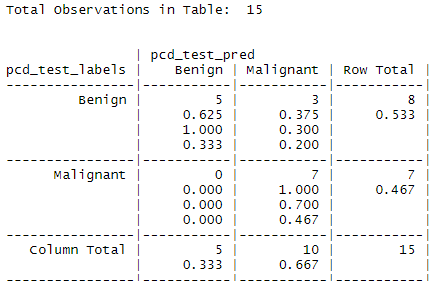
**Working:**

For normalized data, on selecting k=10, 0 false negatives, 3 false positives.

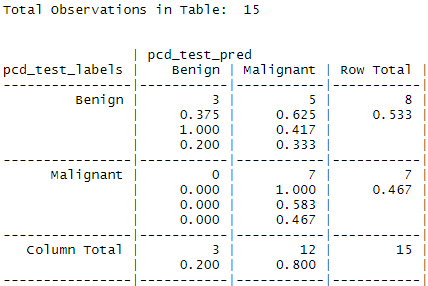


There is misclassification. There are 5 true negative (Actual benign and predicted benign) and 7 true positive (Actual Malignant and predicted Malignant) which were classified correctly. There are no false negatives (Actual malignant predicted benign) and there are 3 observations which are misclassified – false positive (Actual benign and predicted malignant). Let’s scale it.

For standardized data, on selecting k=10, 0 false negatives, 3 false positives. The classifier modification didn’t improve the performance. No difference in performance.

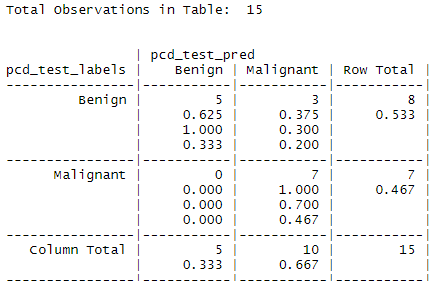


**For normalized data,** **k= 4: 0 false negatives, 5 false positives.**



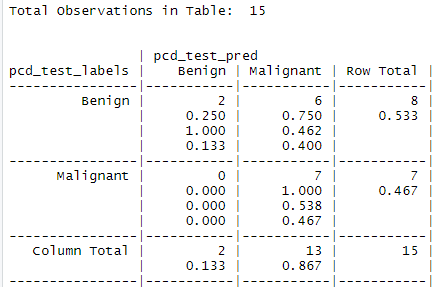
As seen above, there are no false negatives, 5 instances where the tumor was benign but predicted malignant. In healthcare, false negatives should be avoided as much as possible as it may lead to patient believing that he/she is disease free or fit and will not receive the treatment which will pose serious threat to health of these patients and in worst case death.

**For k= 10, 0 false negatives, 3 false positives**



As seen above no instances of false negative which is good which means there were no instances where the tumor was actually malignant but predicted benign. But 3 instances where the tumor was benign but predicted malignant.

**For k= 25, 0 false negatives, 6 false positives.**



As seen above, again there are no instances of false negatives meaning no malignant tumor was predicted benign. There are however 6 instances, where the tumor was benign but predicted malignant.

**Testing alternative values of K**

|  |  |  |  |
| --- | --- | --- | --- |
| k value | False negatives | False positives | Percent classified incorrectly |
| 1 | 0 | 4 | 4 percent |
| 5 | 0 | 5 | 5 percent |
| 11 | 0 | 3 | 3 percent |
| 15 | 0 | 4 | 4 percent |
| 21 | 0 | 5 | 5 percent |
| 27 | 0 | 6 | 6 percent |
| 10 | 0 | 3 | 3 percent |
| 4 | 0 | 5 | 5 percent |
| 25 | 0 | 6 | 6 percent |

**Summary/Conclusion**

The study of the KNN for the prostate cancer example shows the KNN classification approach for the 100 test samples.The results show that there are misclassifications. Classifier was never perfect for any handpicked k value. Higher misclassification for the data on the lower size of K suggests higher misclassification rate, this is due to the data points being grouped closely and being arbitrarily assigned to the neighbor due to conflict or tie. While there were no false negatives, there were false positives for all the k values. Initially there were more false positives till k=10, it dropped near 10,11 and again increased. So for this data, k=10 classifier would be optimal. It is worst in medical industry to classify malignant as benign as it may put the patient’s life in risk.

**References**

Best way to learn kNN Algorithm in R Programming. (2015, March 19). Retrieved from [https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms- programming/](https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-%09programming/)

Lantz, Brett. (2015). Machine Learning with R Second Edition By Retrieved From:

[https://the-eye.eu/public/Books/Programming/Machine%20Learning%20with%20R%20- %20Second%20Edition%20%5BeBook%5D.pdf](https://the-eye.eu/public/Books/Programming/Machine%20Learning%20with%20R%20-%09%20Second%20Edition%20%5BeBook%5D.pdf)

Shrivastava, V. (n.d.). Prostate Cancer. Retrieved from [https://www.kaggle.com/multi8ball/prostate- cancer](https://www.kaggle.com/multi8ball/prostate-%09cancer)