RESEARCH

Introduction to the profile areas of data sciences: project 2

Abstract

Goal of the project: Develop two classifiers to enable the classification of the given class (diagnosis)

Main results of the project: The support vector machine and the logistic regression classifier were accurate in predicting the target groups with an accuracy of about 91% and 90%.

Personal key learnings: We got to know the methods of feature selection and learned how to apply them

Estimated working hours: 8

Project evaluation:

Number of words: 957 Words

1 Scientific Background

In order to diagnose breast cancer, Interactive image processing techniques and linear programming model was applied. Small drop of fluid(sample) was obtained from breast tumor using a small needle. These samples were converted into digital images using a microscope and a camera. Snakes (Active contour model) along with user interaction locates the contours in images. Snake-generated cell nuclei boundaries created in such a way are used to extract different features using computer vision diagnostic system. 10 features with its mean, largest and worst values are yielded after processing 569 images. The challenge was to use these features and find which points can be separated into benign and malignant class. Linear programming model called MSM-T was used to separate the points based on three features namely texture mean, largest area and largest smoothness. Accuracy of 97 percentage was achieved using these three features.

2 Goal

The goal of this project was to perform classification using two different classifiers on Breast Cancer database to diagnose the state of the sample(type M or B). Furthermore, evaluate classifier's accuracy, precision and recall values and perform ROC analysis. Additionally, Compare the results obtained by each individual classifier to analyse which classifier performed better.

3 Data and Preprocessing

A dataset containing 10 different features derived by 567 digital Images extracted by fluid samples from different breast tumor patients was used to diagnose breast cancer. The features are further categorized into its mean,largest and worst values making a total of 30 different features. The database has .data extension and was

read in python using csv reader. Additional and Unnecessary feature such as "ID" was dropped from the database. There exits a "Class" column in the database representing whether the breast cancer is of type M(Malignant) or B(Benign). In order to use this Class to train the model we separated M B and replaced M with 0 and B with 1 using label encoder.

4 Methods

Two classifiers, Support Vector Machine (SVM), which uses a linear kernel algorithm and Logistic Regression (LR) which uses a liblinear algorithm, which follow the same workflow, are implemented by Python. First, normal Python packages and packages needed for statistics are imported. Then we load the csv-data, which for each biopsy sample contains 30 parameters acquired with interactive image processing techniques. After some preprocessing (removal of unnecessary features such as ids), we collected statistical information about the data through summary statistics, distribution histograms, outliers and variable relationships for three attributes (texture mean, area worst and smoothness worst), which were mentioned in the paper[1] as the most promising for tumor classification. Then, we performed feature selection task with correlation, looking for features that are correlated with each other and do not contribute much to explained variance. Then we split the data into train and test sets with the ratio 70:30. Further, we came to the model algorithm based on the training with the train data and evaluate it using test data. Finally, we estimated our model and performed a cross-validation, but before that we had to use the label encoder to get two separate classes. To evaluate the model we compute a confusion matrix, crossvalidation summary and a ROC curve where Sensitivity is a true positive rate (TP/(TP+FN)) and 1-Specificity is a false positive rate (FP/(FP+TN)), so we can get the AUC which is an area value under the ROC curve from each classifier. Then we can compare the models based on this value. The higher AUC, the better the model. To find the most important features we used building functions coef_[0].

5 Results

The statistics results for three attributes show that they are independent from each other, the area worst has outliers and has more separated malignant and benign classes in the distribution histogram than other two attributes. Therefore area worst is a promising candidate for classification of two diagnosis classes. We extracted most important features and found that some attributes have the highest feature importance like concativity_worst, texture_se, concativity_mean and symmetry_mean which have high feature importance for both classifiers (see Figure 1). Area mean which correlates strong with area worst seems not to be significant for classification using SVM or LR as a classification model. Here are the results from our analysis with the test dataset (see Figure 2). The accuracy of SVM and for LR is 91% and 90% (see Figure 3). 0 means benign and 1 means malignant. For the benign class we received a high score for both precision and recall. This means that the classifier provides an accurate result. But for the cancer class the result shows high predictions, but slightly lower recall means that slightly less correct labels are returned, but most of these predicted labels are still correct compared to the training labels. Furthermore, through confusion matrix it is evident that 15 predictions

are false over the sample size of 171 for SVM model. Therefore, probability of error occurring in SVM model can be calculated as [7+8/171 = 0.087] which is 8.87%. Similarly, for the LR model, the probability of error is [13+4/171 = 0.099] which is 9.94%.

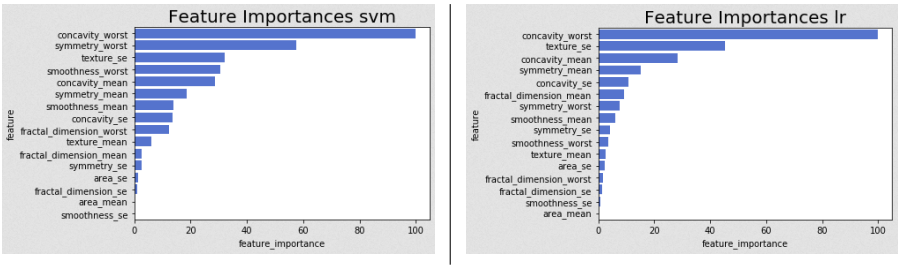

Figure 1 feature importance of SVM and LR

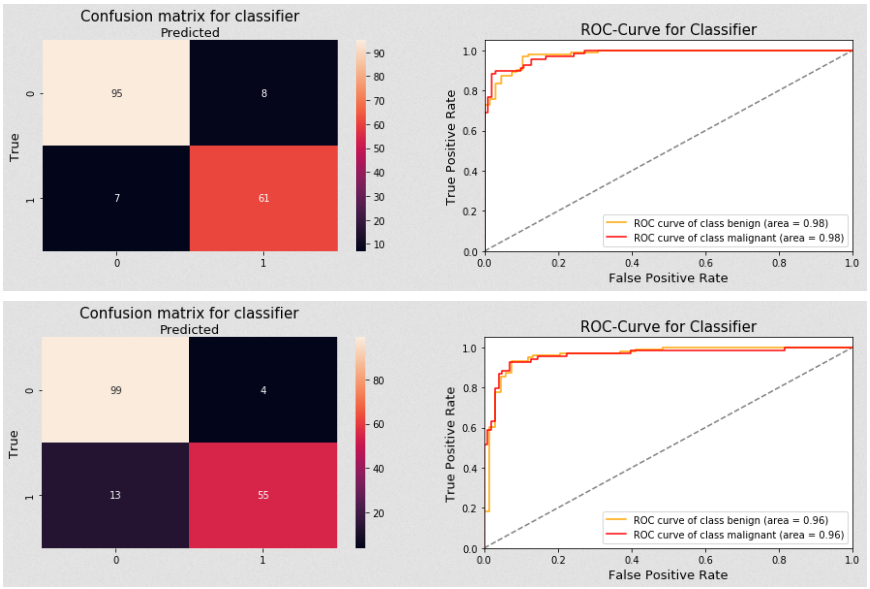

Figure 2 Confusion Matrix and ROC curve of SVM and LR with test set

```
Cross validation accuracy:
        0.9122807017543859

Cross validation classification report

              precision    recall  f1-score   support

           B       0.93      0.92      0.93       103
           M       0.88      0.90      0.89        68

    accuracy                           0.91       171
   macro avg       0.91      0.91      0.91       171
weighted avg       0.91      0.91      0.91       171
```

```
Cross validation accuracy:
        0.9005847953216374

Cross validation classification report

              precision    recall  f1-score   support

           B       0.88      0.96      0.92       103
           M       0.93      0.81      0.87        68

    accuracy                           0.90       171
   macro avg       0.91      0.88      0.89       171
weighted avg       0.90      0.90      0.90       171
```

Figure 3 CrossValidation of SVM and LR with test set

6 Discussion

In this project, the methods for estimating and evaluating the importance of features, filtering the insignificant features to optimise the accuracy of the classifier and making final evaluation show the typical handling of a data classification problem.

References

1. Street, W.N., Wolberg, W.H., Mangasarian, O.L.: Nuclear feature extraction for breast tumor diagnosis. In: Acharya, R.S., Goldgof, D.B. (eds.) Biomedical Image Processing and Biomedical Visualization, vol. 1905, pp. 861–870. SPIE, ??? (1993). doi:10.1117/12.148698. International Society for Optics and Photonics. https://doi.org/10.1117/12.148698