Nika Soltani Tehrani
nsoltanitehraniw@gmail.com

Assignment 1
Breast Cancer Classification

2021-07-30

Assignment description

In this exercise we were facing a binary class classification problem to identify whether one person has breast cancer or not from some information about him like age, BMI, etc.
The data set we worked on was ’Breast Cancer Coimbra Data Set’ to train our model and also tried using Logistic Regression, SVM with linear kernel, SVM with rbf kernel, Decision Tree and KNN algorithms to reach the best accuracy.

1 Data Preparation

In this section, I first checked some primary functions like evaluating the shape of the data set, amount of null values that were zero.
Also I checked whether we are facing an unbalanced data set or not. Classes ’1’ and ’2’ which ’1’ means one person does not have cancer and ’2’ means one person has cancer, had almost the same size so there was no need to balance the data set using functions like SMOTE.
Trying seaborn’s heat map plot showed that almost non of the features have correlation with each other. Also to find those features which have the most impact on the result, I tried seaborn’s pairplot function which it plotted the scatter plot of each two feature. Based on which two features which could separate the classes the best, I realized that features ’Age’ and ’Glucose’ has the most impact on breast cancer.
Based on the diagrams I understood that in average, people who have breast cancer, have glucose more than 100 and insulin resistance score three times more than healthy people as well as we can infer that being under 25 years old has less risk of getting cancer.

2 SVM classifier

Intro

support-vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.
a support-vector machine constructs a hyper plane or set of hyper planes in a high- or infinite-dimensional space, which can be used for classification,

regression, or other tasks like outliers detection.[3] Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.

2.1 SVM classifier using Linear Kernel

Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Data Set.
The result of using linear kernel was getting accuracy of **0.67**

2.2 SVM classifier using RBF Kernel

The radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular
The best result of using RBF kernel was getting accuracy of **0.54**
It has to be said that gamma and C values are key hyper parameterss that can be used to train the most optimal SVM model using RBF kernel. The gamma parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close' Higher value of gamma will mean that radius of influence is limited to only support vectors. This would essentially mean that the model tries and overfit. The model accuracy lowers with the increasing value of gamma. The lower value of gamma will mean that the data points have very high radius of influence. This would also result in model having lower accuracy. It is the intermediate value of gamma which results in a model with optimal accuracy. The C parameter determines how tolerant is the model towards misclassification. Higher value of C will result in model which has very high accuracy but which may fail to generalize. The lower value of C will result in a model with very low accuracy.
In sklearn library the svm model function has an ability to find best gamma and C to get the best result.

2.3 SVM classifier using RBF Kernel with K-fold cross validation

Cross-validation is a re-sampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation. The best result of using RBF kernel with k-fold was getting accuracy of **0.91**

3 Decision Tree

The decision tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for regression problem.
The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.
To evaluate the model, I decided to draw confusion matrix and based on what achieved, I could simply understand metrics like precision, recall and F1 score.
Accuracy
The accuracy score is the fraction of true positives and true negatives over the total number of assigned labels
sum(diagonals in the confusion matrix) / sum (all boxes in the confusion matrix)
Precision
This tells us how many of the values we predicted to be in a certain class are actually in that class. Essentially, this tells us how we performed in terms of false positives.
True positive (number in diagonal)/All positives (column sum)
Recall
This tells us how many of the values in each class were given the correct label, thus telling use how it performed relative to false negatives.
True positive (number in diagonal)/All assignments (row sum)
F1 score
This is a weighted average of precision and recall scale, with 1 being the best and 0 the worst. This uses the harmonic mean, so that the value is closer to the smaller number, and prevents overestimating the performance of the model in cases where one parameter is high and the other low.
2 * (precision * recall)/(precision + recall)
The result of using decision tree was getting accuracy of **0.57**.
Using ”feature importance”method of Decision tree model in sklearn, I got the assurance that ’Age’ and ’Glucose’ have the most impact on the result of the classification.
We can try to improve the model by changing the features used, but we

can also see how it responds to changes in hyper parameters by using GridSearchCV. This performs cross validation on the model by performing the algorithm on multiple runs of the sets of the training set, and tells us how the model responds.
For our purpose, we can change the 'max depth' and 'min samples split' parameters which control how deep the tree goes, and the number of samples required to split an internal node.
Best parameters set found on development set:
'max depth': 4, 'min samples split': 2

4 Logistic Regression

To extract features which have best and impact on the result of the logistic regression model, I tried to use Recursive Feature Elimination (RFE) which is based on the idea to repeatedly construct a model and choose either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features.
This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features but at last the result showed that all the features were important.
The result of using logistic regression was getting accuracy of **0.6**.
Another way I tried to evaluate the performance of the model was using ROC curve. The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.
The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes. The area under the curve became **0.59**.

5 K-NN model

k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.
I tried different k nearest neighbors and the best was 4 with the accuracy of

0.64. This k was where the accuracy on both training and test set was the most see the problem in trade-off.