Machine Learning CSE574: Assignment 3

Priyaman Venu Saranyan (50134222) Tanvi Bhavsar (50133125) Karamveer Choudhary (50134943)
(Group 27)

Results and Analysis

i> Logistic Regression

Accuracies	Train	Test	Validation
LR	92.328	91.92	91.46

Logistic regression is a direct probability. The binary logistic model is used to predict a binary response based on one or more features. Multinomial logistic regression or polytomous logistic regression have more than two categories.

ii> Support Vector Machines

SVM Accuracies	Train	Test	Validation
kernel = 'linear'	97.286	93.78	93.64
kernel='rbf' and gamma=1.0	100	17.14	15.48
kernel='rbf' and gamma='default'	94.294	94.42	94.02

a> Impact of Gamma (γ) and the effect of using a RBF kernel

The linear kernel draws support vectors as straight lines, hence it gets lower accuracy in the Training dataset as the data is non-linear due to the presence of outliers and it the linear kernel is able to provide support vectors upto an accuracy of 97% on the training data and around 93% on the test and validation datasets.

The default value of gamma is set to 1/num_features in sklearn.SVC.svm and the use of a Radial Basis Function kernel helps the classifier to adapt to the non-linearities in the data, so it is more robust to outliers. γ determines the shape of the peak used to raise the data to a higher dimension when using a Gaussian RBF kernel. This can be seen as though the accuracy with respect to the training set is lower compared to its linear counterpart, it achieves higher accuracy with respect to the test and validation datasets.

Using the Radial basis function as a kernel and using a gamma = 1 results in an overfitted scenario and leads to 100% accuracy in the training data set but fails miserably when coming to the test and validation datasets. The width of each bell-shaped surface will be inversely proportional to γ. If this width is smaller than the minimum pair-wise distance for your data, you essentially have overfitting. As the support vectors are exactly (over) fitted to the training data, it is unable to classify the Test and Validation data accurately. This explains the 100% accuracy in the training set but poor performance with respect to the test and the validation data sets.

b> Impact of the Soft Margin Cost (C)

SVM Accuracies	Train	Test	Validation
C = 1	94.294	94.42	94.02
C = 10	97.132	96.1	96.18
C = 20	97.952	96.67	96.9
C = 30	98.372	97.04	97.1
C = 40	98.706	97.19	97.23
C = 50	99.002	97.19	97.31
C = 60	99.196	97.16	97.38
C = 70	99.34	97.26	97.36
C = 80	99.438	97.33	97.39
C = 90	99.542	97.34	97.36
C = 100	99.612	97.4	97.41

C or "Soft Margin" SVM parameter allows some examples to be "ignored" or placed on the wrong side of the margin (the hyperplane).

So larger the margin, more error is allowed. This results in an improvement in the accuracy in the case of a training dataset but results in the hyperplanes being over fitted to the training data set and we can see that the accuracy consistently drops in the test and validation datasets as we keep increasing the value of C