CIS 419/519: Homework 4

{Yupeng Li}

02.26.2020

Although the solutions are entirely my own, I consulted with the following people and sources while working on this homework: **Yifan Yuan, Zhuoyu He**

PART1: PROBELM SET

1 Fitting an SVM by Hand

From the feature vector $\phi(x) = [1, \sqrt{2}x, x^2]^T$
We can expand x_1 and x_2 into 3D space so that they are converted into vectors such that

$$x_1 = [1, 0, 0]^T, x_2 = [1, 2, 2]^T$$

a.) Since the optimal vector w is orthogonal to the decision boundary, it must be parallel to the vector connecting x_1 and x_2.
The vector is then:

$$[1, 2, 2]^T - [1, 0, 0]^T = [0, 2, 2]^T$$

b.) The value of margin is then the L2 norm of the vector connecting x_1 and x_2

$$margin = \sqrt{v_x^2 + v_y^2 + v_z^2} = \sqrt{0 + 4 + 4} = 2\sqrt{2}$$

c.) The L2 norm of vector w can then be determined using the expression:

$$d = \frac{2}{\|w\|_2}$$

Thus,

$$\|w\|_2 = \frac{2}{margin} = \frac{2}{2\sqrt{2}} = \frac{\sqrt{2}}{2}$$

We can then assume w in the form of $[0, 2i, 2i]$ with L_2 norm of $\frac{\sqrt{2}}{2}$
Thus we can get

$$\sqrt{4i^2 + 4i^2} = (\frac{\sqrt{2}}{2})$$

$$2\sqrt{2}i = \frac{\sqrt{2}}{2}$$

$$2i = \frac{1}{2}$$

Thus,

$$w = [0, \frac{1}{2}, \frac{1}{2}]$$

d.) From the fact that

$$y_1(w^T\phi(x_1) + w_0) \geq 1$$

and

$$y_2(w^T\phi(x_1) + w_0) \geq 1$$

and

$$y_1 = -1, y_2 = 1$$

We know that

$$(w^T\phi(x_1) + w_0) \leq -1$$

and

$$(w^T\phi(x_2) + w_0) \geq 1$$

Since $w^T\phi(x_1) = 0$, $w_0 \leq -1$ and since $w^T\phi(x_2) = 2, w_0 \geq 1 - 2$, Thus,

$$-1 \leq w_0 \leq -1$$

$$w_0 = -1$$

e.) Thus the discriminant can be expressed as in the following equation:

$$h(x) = -1 + \frac{\sqrt{2}x}{2} + \frac{x^2}{2} \tag{1}$$

2 Support Vectors

The size of the margin should either increase or stays the same for the dataset. This is because once you remove the support vector, the margin would either stay the same because of a support vector of same length or expand due to the shortest vector has been removed.

3 Challenge: Generalizing to Unseen Data

3.1 The Challenge Data

Some features are removed from the challenge data including data of entry, countries funded and IDs since these do not help with classifying the dataset. Other categorial features have been One-Hot-Encoded.

3.2 The Boosted Decision Tree Classifier

The BoostedDT classifier is submitted in the .py file. Based on cross validation, the BoostedDT has highest accuracy when the number of iterations is 100 and the tree depth is 10. This gives the highest CV score of around 77% and train accuracy of approximately 82.7%

3.3 Comparing Boosted Decision Trees with SVMs

Using SVM with Gaussian Kernel and one-versus-one training strategy, the model gives a training accuracy of 83.6% and a CV score of approximately 76.2%

3.4 Training the Best Classifier

The best classifier I discovered during the training process is the DecisionForest Classifier by Sklearn. In the process of training, I gradually increased the tree depth until the margin is diminishing. The result training accuracy is about 85% and the result CV score is 80.6%

3.5 Comparing Your Algorithm

Preprocessing

The dataset is preprocessed using the process below: The dataset is first merged with the training labels so that the instances with nan predictions will be eliminated. After that, features that are not useful or redundant including id, date of entry and location are removed from the dataset. Categorial features are then One-Hot-Encoded except for country-funded feature since even it is a categorial feature, it has more than 10000 categories and it doesn't help much with classifying. After OHE, the missing features are then imputed using the simple-Imputer from Sklearn library. For the SVM classifier, an extra standardization step is performed on the dataset.

The Best Classifier

The best classifier I discovered during the training process is the DecisionForest Classifier by Sklearn. In the process of training, I gradually increased the tree depth until the margin is diminishing. The result training accuracy is about 85% and the result CV score is 80.6%

Result and Discussion

Method	CVScore	Training Score
BoostedDT	0.77	0.827
SVC	0.762	0.836
DecisionForest	0.806	0.85

The table above compares my results using different training models. From above we can see that the Decision Forest classifier has the highest CVScore and followed by BoostedDecision Tree and SupportVector Machine using Gaussian. For the Decision Forest model, it reaches its highest generalization accuracy at depth of 20. For the SVC model, it reaches its highest accuracy when using the rbf Kernel and when the data is standardized. For the Boosted DT model, it reaches its highest accuracy when the maximum depth is 10.