CIS 419/519: Homework 4

Jiatong Sun

02/29/2020

Although the solutions are entirely my own, I consulted with the following people and sources while working on this homework: *Jing Zhao*
$https://oeis.org/wiki/List_of_LaTeX_mathematical_symbols$
$https://tex.stackexchange.com/questions/122778/left-brace-including-several-lines-in-eqnarray$
$https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html$
$https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html$

1 Fitting an SVM by Hand

a. As is given in the problem:

$$x1 = 0, \quad x2 = \sqrt{2} \tag{1}$$

and

$$\phi(x) = [1, \sqrt{2}, x^2]^T \tag{2}$$

We know that

$$\phi(x_1) = [1, 0, 0]^T \quad \phi(x_2) = [1, 2, 2]^T \tag{3}$$

Since the optimal vector $\boldsymbol{w}$ is orthogonal to the dicision to the decision boundary, it is parallel to the vector connecting $\phi(x_1)$ and $\phi(x_2)$.

Since

$$\phi(x_2) - \phi(x_1) = [1, 2, 2]^T - [1, 0, 0]^T = [0, 2, 2]^T \tag{4}$$

So $[0, 2, 2]^T$ is a vector that is parallel to the optimal vector $\boldsymbol{w}$.

b. The margin is the distance between the two points in the 3D space.

$$margin = ||\phi(x_2) - \phi(x_1)|| = \sqrt{(1-1)^2 + (2-0)^2 + (2-0)^2} = 2\sqrt{2} \tag{5}$$

c. From the result of a, we can assume that

$$\boldsymbol{w} = [0, 2t, 2t]^T \tag{6}$$

So

$$||\boldsymbol{w}|| = \sqrt{0^2 + (2t)^2 + (2t)^t} = \sqrt{8t^2} = 2\sqrt{2}t \tag{7}$$

According to the relationship between $||w||$ and the length of the margin, we know that

$$d = \frac{2}{||w||} = \frac{1}{\sqrt{2}t} = 2\sqrt{2} \tag{8}$$

or

$$t = \frac{1}{4} \tag{9}$$

So

$$\boldsymbol{w} = [0, \frac{1}{2}, \frac{1}{2}]^T \tag{10}$$

d. According to SVM requirement,

$$\begin{cases} y_1(\boldsymbol{w}^T\phi(x_1)+w_0) \geqslant 1 \\ y_2(\boldsymbol{w}^T\phi(x_2)+w_0) \geqslant 1 \end{cases} \tag{11}$$

or

$$\begin{cases} -1\times(0+w_0) \geqslant 1 \\ 1\times(2+w_0) \geqslant 1 \end{cases} \tag{12}$$

$$-1 \leqslant w_0 \leqslant -1 \tag{13}$$

So

$$w_0 = -1 \tag{14}$$

e.

$$h(x) = \boldsymbol{w}^T\phi(x) + w_0 = \frac{x^2}{2} + \frac{\sqrt{2}x}{2} - 1 \tag{15}$$

2 Support Vector

There are two possibilies:

1. Size of maximum margin increases, if a support vector determining the shortest margin is removed.

2. Size of maximum margin stays the same, if the removed vector is not the one determining the shortest margin.

3 Challenge: Generalizing to Unseen Data

Preprocessing

1. Sort the data in X and y in ascending direction and eliminate rows whose id only appears in X or y. After this operation, X and y are aligned by an ascending id number.

2. Drop useless features. I dropped the features below: ['id', 'Date of entry', 'Country funded by', 'oompa loomper', 'Region code', 'District code', 'Chocolate consumers in town', 'Does factory offer tours', 'Recorded by','Oompa loompa management', 'Payment scheme', 'management group'].

3. Drop columns whose missing ratio is higher than 50%.

4. Define categorical features (even though some features look like numerical features, if they are defined as categorical in the pdf, they need to be converted to object dtypes).

5. Fill missing data (mean for numerical and mode for categorical). Use OHE to process the categorical data.

6. Drop the id column of label.

7. For unlabeled data set, after step (2)-(5), we still need to add missing features and reduce redundant features with respect to the training data set or otherwise the data cannot be predicted.

The Best Classifier

The best classifier I found is the Support Vector Machine (SVM). Before training the model, I standardized the training data set, which turns out to be a great way of increasing the accuracy. All the parameters are set as default in according to skicit.

Results and Discussion

Algorithms Comparison		
	training accuracy	generalized accuracy
Boosted Decision Tree	0.7186	0.7143
Support Vector Machine	0.7628	0.7429
Logistic Regression	0.7196	0.7189

According to the result table, we can notice the support vector machine generates a better performance than the other two algorithms. But this result is also based on the fact that during the training process, SVM took much more time than the other two, so we can simply arrive at the conclusion that in this scenario, SVM generates a more complicated model. The boosted Decision Tree and Logistic Regression may have perfomance gain if we tune the parameters and make the model a bit more complex, such as increasing the maximum iteration number or maximum tree depth for BoostedDT.