CIS 419/519: Homework 4

{Yongxin Guo}

Although the solutions are entirely my own, I consulted with the following people and sources while working on this homework: Jiangzhu Heng, Yihang Xu

1 Part I: Problem Set

1.1 Fitting an SVM by Hand

a). Since we are looking for a vector that is orthogonal to the decision boundary, then it can be directly computed by taking the difference between two 3D points to get the directional boundary:

$$\phi(x_1) = [1, 0, 0]^T$$

$$\phi(x_2) = [1, 2, 2]^T$$

$$v = \phi(x_2) - \phi(x_1) = [\mathbf{0}, \mathbf{2}, \mathbf{2}]^{\mathbf{T}}$$

b). The margin is defined by the distance between these 2 3D points:

$$margin = ||\phi(x_2) - \phi(x_1)||_2 = 2\sqrt{2}$$

c). Using the equation provided:

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

And we know the unit vector of w can be computed using results from a):

$$unit(w) = \frac{v}{||v||} = \frac{[0, 2, 2]^T}{2\sqrt{2}} = [0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]^T$$

Then,

$$w = unit(w)||w||_2 = [0, \frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]^T \times \frac{\sqrt{2}}{2} = [\mathbf{0}, \mathbf{0.5}, \mathbf{0.5}]$$

d). Using the constraints (2)2-(3):

$$-1 \times (0 + w_0) \geq 1$$

$$1 \times (2 + w_0) \geq 1$$

So,

$$w_0 \leq -1 \; and \; w_0 \geq -1$$

Then,

$$\mathbf{w_0 = -1}$$

e).

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

1.2 Support Vectors

If removing one support vector, the maximum margin may increase or stay the same because there may exists one or multiple data point along either one support vector line. Thus by removing one support vector, the maximum margin may stay the same if there are multiple support vectors at one side. The maximum margin can increase if there is only one support vector

2 PART II: PROGRAMMING EXERCISES

2.1 Comparing Algorithms

a). Pre-processing: The pre-processing procedure is outlined as the following:

- **Removing unnecessary information by reasoning:** 1) **Recorded by**, which is done by one exact same person, so it doesn't provide potential value to the model. 2) **Date of entry**, it will not have noticeable impact on the label by reasoning. 3) **id**, it will not have noticeable impact on the label by reasoning. 4) **Region and District Code**, redundant information as there is already longitude, lattitude and Location that represent the geographic information. 5) **Country of Factory**, speculated to be not relevant to the label and it also has too many data points that makes it impossible to do OHE
- **Missing Value Handling:** Country funded by, oompa loomper, Does factory offer tours, Oompa loompa management, Official or Unofficial pipe have missing values, all the missing values are replaced by its separate modes since they are all categorical features
- **One-Hot Encoding (OHE):** One-hot encodeing all the rest of categorical features
- **Outlier Handling:** By observation, all the numerical values have a lot of zeros, which will have effect on the models. Thus, all the zeros are replaced by the mean values of that numerical features (zeros are not counted during the mean finding process)
- **Consistence between Training and Test Data:** The features in both data set are extracted by finding the intersections of both

b). **The best classifier:** The best classifier is the adaboost-SAMME with the decision tree as the base learner. The numOfIterations is set to **100, and the max depth of the tree is set to 13**. The train data has to go through the pre-processing steps as described in a) and fit it into the adaboost-SAMME, then the model is well-trained so that it can be used to predict the unseen data that also has to go through the identical pre-processing steps.
c). Parameters:

	Ada-SAMME	SVC	Single Decision Tree
Training Accuracy	0.993	0.793	0.75
Generalization Accuracy	0.803	0.786	0.7

Table 1: Performance Comparison

Ada-SAMME: Depth = 13, Iteration number = 100
Single Decision Tree: ccp = 0.02
Discussion: It is found that the training accuracy correctly matches with the trend of the generalization accuracy with Ada-SAMME being the highest for both performance matrices. For the best classifier, Ada-SAMME, it is found that by increasing the number of iterations above 100, the accuracy won't be improved at all. However, the depth of the tree plays a big role in the performance, which is found that the accuracy is the highest when the depth is 13 and it even drops when the depth is higher.