**SVM – Classification Problem**

The purpose of this assignment to examine what happens in a classification problem when the class attributes overlap. To begin the assignment, the experiment was recreated, and the predictions were displayed using a bar chart and scatterplot. After, the SVM model was trained to use the 1,000 data length to predict the 10,000 data length. For the final question, the subject of ambiguous data was discussed in-depth.

**Questions:**

1. **Recreate the R part of this experiment using your computer.**

The screenshots below recreate the R portion of the experiment. To begin, I loaded the library “e1071” and then uploading the data sets srd1 and srd2. Afterward, I created a vector of 1,000 zeros for y1 and a vector of 1,000 ones for y2. These vectors will help with the results for srd1 and srd2. Next, y was created by concatenating y1 and y2. Then x was created to develop a 2,000 row by 2 column matrix using row binding. Afterward, the SVM model was created, and the predict function was used to generate the predictions. For the predictions, it is expected that we will have zeros for the first 1,000 values and ones for the last 1,000 values. However, in the summary, we see that the model predicted 883 zero’s and 1,117 ones. The final step plots the results with the function plot(array(p)).

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. **See if you can find other ways to display the predictions.**
2. One way to display the predictions is with a bar chart. A simple bar chart was created with the command plot(p).

A picture containing screenshot

Description automatically generated

1. Another way to display the predictions is with a scatter plot. The scatterplot was created with the following commands:

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. A better way to display the predictions is with a scatter plot showing the ambiguity data points. This scatterplot was created with the following commands:

A picture containing screenshot

Description automatically generated

A close up of a map

Description automatically generated

1. **Train the SVM model using the 1,000 length data and then predict the 10,000 length data.**

For this question, the goal is to train the SVM model using the 1,000 length data to predict the 10,000 length data. To start the process, I imported the rd1 and rd2 files and then created x2 by row binding the 10,000 data set. Afterward, I used the prediction function. The prediction function uses the model that was built in the previous steps using sd1 and srd2, and x2 to predict the 10,000 length data. For the predictions, we anticipate the first 10,000 to be zeros and the second 10,000 to be ones. However, when we test the predictions, the model predicts 8,745 zeros and 11,255 ones. The final step plots the results.

While training the SVM model, it is crucial to limit problems such as overfitting. One solution is to use cross-validation. The goal of cross-validation is to use the first training data to generate multiple mini train-test splits. These splits are used to tune the model (EliteDataScience, 2017).

A screenshot of a cell phone

Description automatically generated



A picture containing screenshot

Description automatically generated

1. **Is there anyway to remove the ambiguity between the two classes given the existing data?**

A challenge is machine learning is the classification of datasets with ambiguous data. Ambiguous data is the outcome of overlapping feature values. In this instance, there is a bound on classification performance. According to Hashemi and Trappenberg (2002), separating ambiguous data can considerably improve the performance of the classification process. In their research, they purpose a new approach where ambiguous data is separated from the data that are much easier to classify to prevent their influence on the classification process (Hashemi & Trappenberg, 2002). With this approach, there is some data loss, but the classification improves significantly.

The following explains Hashemi and Trappenberg's (2002) separation scheme. This approach can be used for both outliers and ambiguous data. Hashemi and Trappenberg (2002) state the following:

On training data:  
1. Train classifier 1 using all of the training data.  
2. Use the information from classifier 1 to divide all points into 2 classes. For instance, A (typical) and B (atypical).   
3. Train classifier 2 on the training data with new labels (A and B); classifier 2 is atypical detector (separator).  
4. Train an additional classifier 3 on only A (typical data) using their original labels.

On test data:

5. Use classifier 2 to remove potential atypical data from the test set (cleaning test data): 2 classes A1 and B1.

6. Use classifier 3 for the classification of A1 data. Use original labels to calculate the performance measures (Hashemi & Trappenberg, 2002).

**What about if you could add additional measurements for each data point.**

According to Hashemi and Trappenberg (2002), bounded support vectors (BSVs) could also be used to separate atypical points from the typical ones when ambiguous data is due to overlapping feature values. Alternatively, CP curves can also be used to show if ambiguous data are an issue and might limit the performance of the classification process (Hashemi & Trappenberg, 2002).

**Reference**

EliteDataScience. (2017, September 7). Overfitting in Machine Learning: What It Is and How to Prevent It. Retrieved from https://elitedatascience.com/overfitting-in-machine-learning

Hashemi, S., & Trappenberg, T. (2002). Using SVM for Classification in Datasets with Ambiguous Data. *SCI*.