Module 3 Technique Practice
Suport Vector Machine

ALY 6015

Jeff Hackmeister

2025-04-26

Introduction

Working with the Polycystic Ovary Syndrome PCOS dataset from Kaggle, we will be using a support vector machine (SVM) for a classification exercise to model the diagnosis of Polycystic Ovary Syndrome (PCOS), a prevalent but severely under diagosed hormonal disorder that can significantly impact women's health and quality of life. This data was collected from patients seen at 10 different hospitals across Kerala, India.

To begin, we'll load our needed libraries and dataset. While the data originally came from Kaggle, I am using a cleaned version from a previous project from the dataset that I have loaded to GitHub.

```
library(e1071)
library(arules)
library(caret)
library(tidyverse)
data <- read.csv(
  "https://raw.githubusercontent.com/jhackmeister/ALY6040/refs/heads/main/Final%20Project/fi
```

To confirm our data, we'll use the str function.

```
str(data)
```

```
'data.frame':   3000 obs. of  7 variables:
 $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Age      : int  29 20 23 19 19 23 21 26 25 31 ...
 $ BMI      : num  21.2 20.5 23.1 32.7 25.9 20.6 25.3 25.2 25.7 29 ...
 $ Men_Irrg : chr  "No" "No" "No" "Yes" ...
 $ T_Level  : num  46.1 59.4 69.3 77.7 49.4 36.7 44.2 32.6 96.8 58.3 ...
 $ AC_Count : int  9 6 10 37 5 5 4 4 37 6 ...
 $ PCOS_diag: chr  "No" "No" "No" "Yes" ...
```

And we can see that we have 3,000 observations of 7 variables. Our variables are

- X - this a row count number and is not needed for this study, we'll drop it
- Age - the age of the participant in the study
- BMI - body mass index is the ratio of body weight to height
- Men_Irrg - an indication of menstrual irregularities, this should be converted to a factor datatype
- T-Level - the patient's measured testosterone levels

- AC_Count - count of patient's antral follicles
- PCOS_diag - indication of a positive or negative PCOS diagnosis, this should also be converted to a factor

We'll perform the identified cleaning procedures and look at the first few lines of the dataset.

```
data <- data %>% select(-X)
data$PCOS_diag <- as.factor(data$PCOS_diag)
data$Men_Irrg <- as.factor(data$Men_Irrg)
head(data)
```

```
  Age  BMI Men_Irrg T_Level AC_Count PCOS_diag
1  29 21.2       No    46.1        9        No
2  20 20.5       No    59.4        6        No
3  23 23.1       No    69.3       10        No
4  19 32.7      Yes    77.7       37       Yes
5  19 25.9       No    49.4        5        No
6  23 20.6       No    36.7        5        No
```

Running the model

Next, we will use the svm function from the e1071 library to build an SVM model for our dataset.

```
model <- svm(PCOS_diag ~ ., data = data)
```

With the model created, we can review the model using the summary function.

```
summary(model)
```

```

Call:
svm(formula = PCOS_diag ~ ., data = data)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
```

```
Number of Support Vectors:  64

 ( 32 32 )


Number of Classes:  2

Levels:
 No Yes
```

```
head(model$index)
```

```
[1]  26 212 264 295 360 477
```

The C-classification indicates that we are using classification, rather than regression, SVM. The radial kernel indicates that the model will incorporate non-linear relationships between the provided variables. The cost 1 is the default for the function and the 2 classes indicate our two outcomes, a Yes or No for a PCOS diagnosis. The even split between support vectors, 32 for both yes and no, is a indications of a clear delineation in the data.

Prediction

With the data fit to the model, we will move to making predictions from the model. To do so, we'll use the predict function from the arules library.

```
predictions <- predict(model, data)

head(predictions)
```

```
  1   2   3   4   5   6
 No  No  No Yes  No  No
Levels: No Yes
```

```
# Confusion Matrix
table(Actual = data$PCOS_diag, Predicted = predictions)
```

```
      Predicted
Actual   No  Yes
   No  2399    1
   Yes    2  598
```

The head function shows the predictions made for the first six observations, predicting no for all observations except number 4.

The confusion matrix shows the summarized results for the entire data set. There were 2,399 true negatives (the negative prediction was correct) with 1 false negative. There were also 598 true positive (a correct positive prediction) with 2 false positives. These are very strong results for our model.

```
(2399 +598) / (3000)
```

```
[1] 0.999
```

Train/Test

To further examine the performance of the model, we'll split the dataset in test and train sets, using 80% of the data for training and leaving 20% for testing. Next, we'll again use svm to fit a model to the training data and then use the resulting model to make prediction from the held over test data.

```
set.seed(123)
train_index <- createDataPartition(data$PCOS_diag, p = 0.8, list = FALSE)
train_data <- data[train_index, ]
test_data <- data[-train_index, ]

model_train <- svm(PCOS_diag ~ ., data = train_data)
test_pred <- predict(model_train, test_data)
confusionMatrix(test_pred, test_data$PCOS_diag)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  479   0
       Yes   1 120

               Accuracy : 0.9983
                 95% CI : (0.9907, 1)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9948
```

```
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9979
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9917
             Prevalence : 0.8000
         Detection Rate : 0.7983
   Detection Prevalence : 0.7983
      Balanced Accuracy : 0.9990

       'Positive' Class : No
```

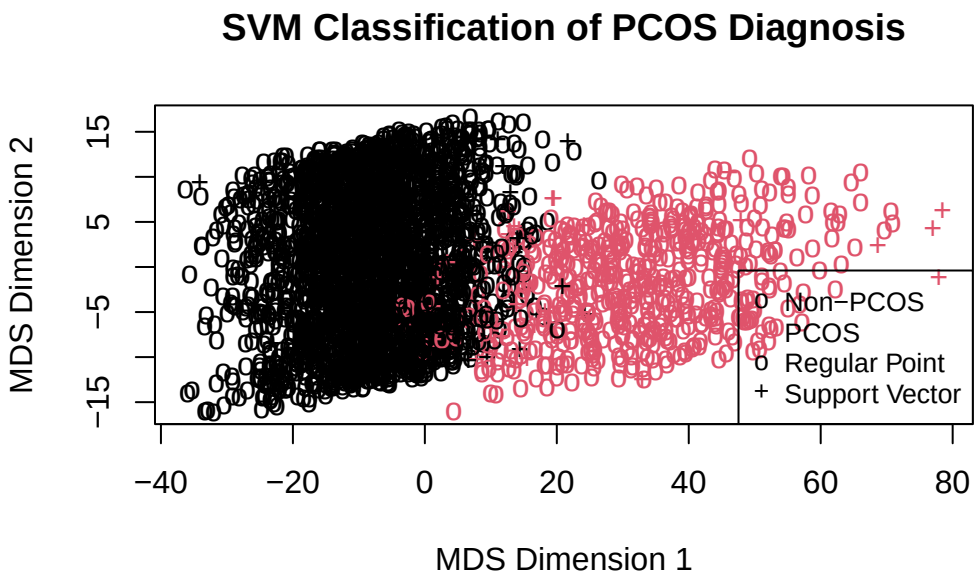
Once again, we see very strong results for the model, with 99.83% accuracy and a 95% confidence interval of 99.07 to 100%. There were also 0 false negatives in this test - which is very important for this particular subject matter.

Visualization

Next, we will use the dist function from arules to create a distance matrix. This measures the distance between the data points in a multidimensional space. To do so, we must first limit the data to only numerical values. Once we have those distances calculated, we'll use the cmdscale function to convert those values into 2D coordinates that we can plot.

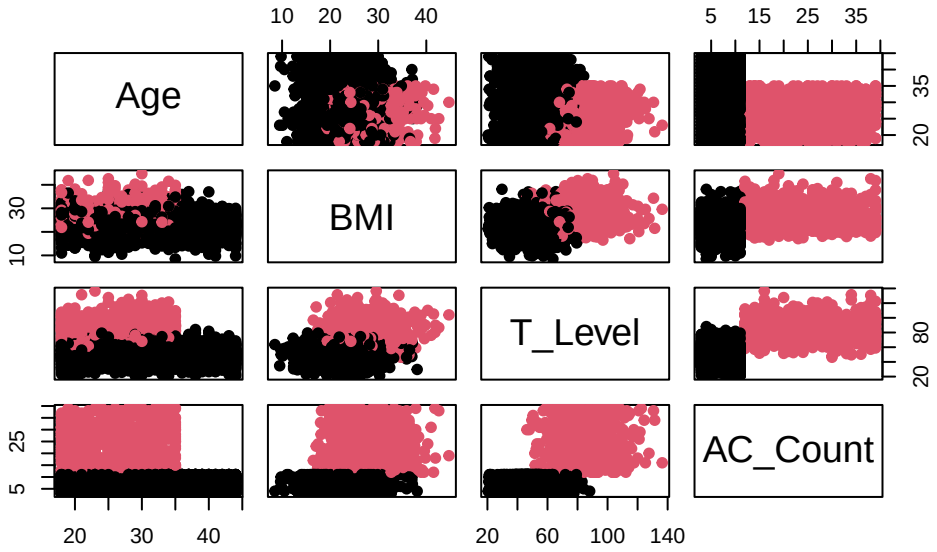
```r
# Extract only numeric columns for distance calculation
numeric_features <- data[, sapply(data, is.numeric)]

# Create distance matrix
dist_matrix <- dist(numeric_features)

# Apply MDS to get 2D coordinates
mds_coords <- cmdscale(dist_matrix)

# Create the plot
plot(mds_coords,
     col = as.integer(data$PCOS_diag),
     pch = c("o", "+")[1:nrow(data) %in% model$index + 1],
     main = "SVM Classification of PCOS Diagnosis",
     xlab = "MDS Dimension 1",
     ylab = "MDS Dimension 2")
```

```
legend("bottomright",
       legend = c("Non-PCOS", "PCOS", "Regular Point", "Support Vector"),
       col = c(1, 2, 1, 1),
       pch = c("o", "o", "o", "+"),
       cex = 0.8)
```



The resulting plot shows the distribution of the observations from the dataset. Positive diagnoses are in red, while negatives are in black. The 64 support vectors identified in the model are indicated by a + while the other observations are shown with a o. Despite the strong results we saw in the model above, there is not as clean of a delineiation in the data from this plot. To look for a cleaner separation, we'll look at each of the variable pairs independently.

```
pairs(numeric_features, col = as.integer(data$PCOS_diag), pch = 19)
```



This gives us a much better view into which variable combination provides the best distinction in diagnosis outcomes. Age has the weakest predictive influence, while BMI, testosterone and AC count produce much cleaner results.

Conclusion and Next Steps

The results of our SVM model were very strong, indicating that these variables are highly useful in predicting a PCOS diagnosis. Age appears to be of limited use in the model, which makes some logical sense given the relatively narrow age band included in a study of reproductive adult women, 18 to 44 in this dataset. Ideally, we could use this model on a new dataset collected from a different geographical area to test the effectiveness on different demographic samples. It would also be quite helpful to add additional demographic variables such as race, socioeconomic status, and access to regular healthcare as they tend to correlate strongly with many other health outcomes.

References

[1] *arules package | R Documentation*. (2012). Rdocumentation.org.

https://www.rdocumentation.org/packages/arules/versions/1.6-4

[2] *Classifying data using Support Vector Machines(SVMs) in R*. (2018, August 28). GeeksforGeeks.

https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/

[3] *cmdscale function - RDocumentation*. (n.d.). Www.rdocumentation.org.

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cmdscale

[4] *e1071 package - RDocumentation*. (2024). Rdocumentation.org.

https://www.rdocumentation.org/packages/e1071/versions/1.7-16

[5]Kottarathil, P. (2020, July 11). Polycystic ovary syndrome (PCOS). Kaggle.

https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

[6]Tierney, N. (2025, April 9). *Quarto for Scientists*. Njtierney.com. https://qmd4sci.njtierney.com/

[7]World Health Organization. (2025, February 7). *Polycystic ovary syndrome*. World Health Organization; World Health Organization.

https://www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome