CLASSIFICATION METHODS FOR SUPPORT VECTOR MACHINES

JULIA ANDRONOWITZ
B.S., Mathematics

May 2023

Abstract

The purpose of this thesis is to give an introduction to the concept of Support Vector Machines in Machine Learning. We will first outline the idea of classification, including the maximal margin classifier and the support vector classifier. Examples of each will be given using programming languages such as R and Python. Then, we will move onto support vector machines and the use of kernels with example data. We will implement the techniques previously described in a real data set and finish by discussing applications of SVMs and examining the documentation of the support vector machine modules in R and Python.

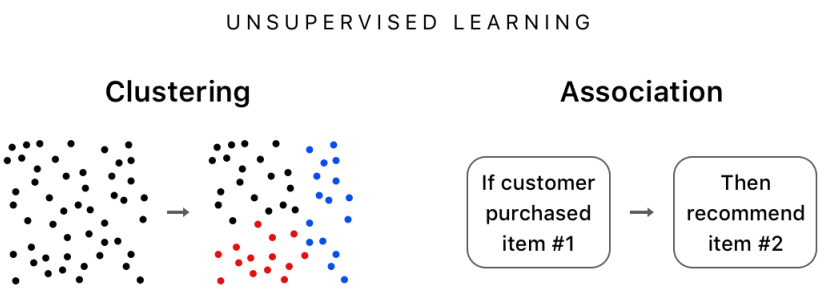
University of Connecticut
Department of Mathematics
Advisor: Dr. Jeremy Teitelbaum

1 INTRODUCTION

The first discoveries regarding machine learning date back to the 1950s. Known as the "Turing Test", mathematician Alan Turing attempted to discern if a computer could fool a human into thinking it is also a human **?**. In the years to follow, computer codes turned into learning programs that evolved the more times the code was run. In 1957, the first neural network was created which simulated a human brain's thought process. Ten years later, a basic pattern recognition algorithm was created called the "nearest neighbor" algorithm. Come the 1990's, accessibility to computers and advances in computers has exponentially increased. Scientists begin using a data-driven approach as large amounts of data are available, which allows computer algorithms to analyze data and learn from their results. Fast forward a few decades and we now have deep learning, survival analysis, and unsupervised learning.

Machine learning algorithms fall into one of two categories: supervised or unsupervised learning. Essentially, unsupervised learning arises when each observation does not have an associated response. This allows the machine to find certain patterns among the data and draw conclusions. Types of unsupervised learning include clustering and association. Clustering problems involve grouping the data based on the features. It is unknown how many groups are present in the data set when we begin, and so we rely on the machine to distinguish between groups. For example, we may want to group different animals based on their hunting/gathering methods and what they consume. Perhaps some animals tend to graze, others tend to scavenge, and others hunt. The algorithm would try to split the data into these three groups. In contrast, association rule learning problems focus on generalized trends between the groups. These types of trends are widely applicable to large portions of the data. For example, if one animal tends to graze on grass and other plants, the machine might suggest that another animal who grazes on grass will also eat other plants. Another common example is that people who buy one product are likely to buy another product **?**.

In contrast, supervised learning uses data that is well-labeled to teach our model and either infer or predict. Inference problems aim for a better understanding of the relationship between the response variable and the features. On the other hand, prediction problems relate to developing a model that accurately fits the response variable to the predictors. Each observation in a data set has associated predictors and a response variable. Predictors are the input variables and can go by a variety of names such as variables, independent variables or features **?**. Predictors usually



4801400

Source: Data Driven Investor

correspond to the variables $x_1, x_2, ..., x_n$ where each x_n is an input variable. The response variable is the dependent variable or what we are trying to measure, usually denoted as y. Let's take a look at some penguin data, shown in Table **??**.

Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Delta 15 N (o/oo)	Delta 13 C (o/oo)	Species
50.2	18.7	198	3775	9.39305	-24.25255	Gentoo
39.5	17.4	186	3800	8.94956	-24.69454	Adelie
44.9	13.8	212	4750	8.11238	-26.20372	Chinstrap
52.2	17.1	228	5400	8.36701	-25.89834	Chinstrap
50.8	19	210	4100	9.98044	-24.68741	Gentoo
42.5	20.7	197	4500	8.67538	-25.13993	Adelie

Table 1: Penguin Data

In this example, the culmen length, culmen depth, flipper length, body mass, Delta 15 N and Delta 13 C are the predictors. Based on these variables, we want the model to predict which species the penguin is. So, each body measurement is a predictor and the species is the response variable. In supervised learning, we often have training and testing groups for our data. Training and test groups are a common practice in machine learning to both classify existing data points and assess the underlying accuracy of the model on known data. Training data is data used to teach our model how to estimate the response variable. Once the model is trained, we can apply the algorithm to the testing data which the model has not previously seen. Since we already have the response variable for this set in the original data, we can compare

the model's predicted y-values, denoted $\hat{y}$, with the actual y-values. In machine learning, it is common to see the data split into about 80% training and 20% testing data, but these values can be manually specified in the code **?**. Suppose we collect the data for the ten students at the end of the year. Then, we have all their test scores for the year. We take 8 students and train the model based on these observations. Then, we use the remaining two students as the test set.

Supervised learning can be categorized into either regression or classification problems. Generally, this has to do with the fact that data is either qualitative (categorical) or quantitative (numerical). Regression typically uses a quantitative response variable. In regression, we aim to create a model that uses features to accurately predict the response variable. We may want to see how a population's access to clean drinking water impacts their average life expectancy. Linear regression can be used in this case to see whether a correlation between the two variables exist and how strong that correlation is. We can also look at how accurate the model is at predicting the response variable as well as investigate whether the relationship is linear or non-linear. On the other hand, classification is common with a quantitative response variable. In these types of problems, the data is grouped into specific categories based on the given features. The model then is able to predict the category of an unknown data point. Similarly, we can also use tools to predict the accuracy of the model and determine if the relationship is linear or non-linear.

Developed in the 1990s, support vector machines are an approach to classification problems now widely used among data scientists. In the following paper, we will discuss what a support vector machine is, how one can be implemented among a data set and practical applications of SVMs.

2 HYPERPLANES

We first begin by defining a hyperplane. Specifically, a hyperplane is "a flat affine subspace of dimension $p - 1$" in a p-dimensional space **?**. In the two-dimensional space, a hyperplane looks like a straight line bisecting the data points into two distinct groups. We call this data linearly separable, as the data can be separated into two distinct groups by the hyperplane. In three dimensions, we see the formation of a plane. In higher dimensions, it can be hard to visualize the hyperplane, but the concept still applies.

Mathematically, a hyperplane in $\mathbb{R}^2$ is the equation $0 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Note that this is the equation of a one-degree polynomial, or a straight line. In higher dimensional spaces, a hyperplane has the form

$$0 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

We can generalize the equation of a hyperplane to be

$$f(x) = \beta_0 + \beta \cdot x$$

where $f(x) = 0$ and β is a non-zero vector $\beta = (\beta_1, \beta_2, ..., \beta_p)$ in $\mathbb{R}^p$ in which the dot product of β and x is computed. For any $x = (x_1, x_2)^T$ where the equation holds, we say the point is on the hyperplane. However, points that do not lie on the hyperplane will have

$$0 \neq \beta_0 + \beta \cdot x$$

Moreover, the points that have $0 > \beta_0 + \beta \cdot x$ will lie on one side of the hyperplane while the points $0 < \beta_0 + \beta \cdot x$ will lie on the other. In this, we can see the concept of linearly separability. In a data set that is linearly separable and each observation is in one of two distinct groups, the hyperplane will bisect the data points in such a way that every point in one group is greater than zero while every point in the alternate group is less than zero.

To examine a hyperplane in two dimensions, we generate two distinct groups of data using *R*.

We first set $n = 200$. This will be the length of our data set. Then, we use the function **rnorm** to generate a set of data points from the normal distribution with mean 1 and standard deviation 0.3. The matrix function transforms the 400 randomly generated points into a [200 x 2] matrix. The same is done for a second group, instead with mean -1 and standard deviation 0.3. The **rbind** function combines these two groups into one vector, namely x.

Now, we plot the function. The **abline** functions add lines at $x = 0$ and $y = 0$ as well as add one such separating hyperplane. Figure **??** shows the output.

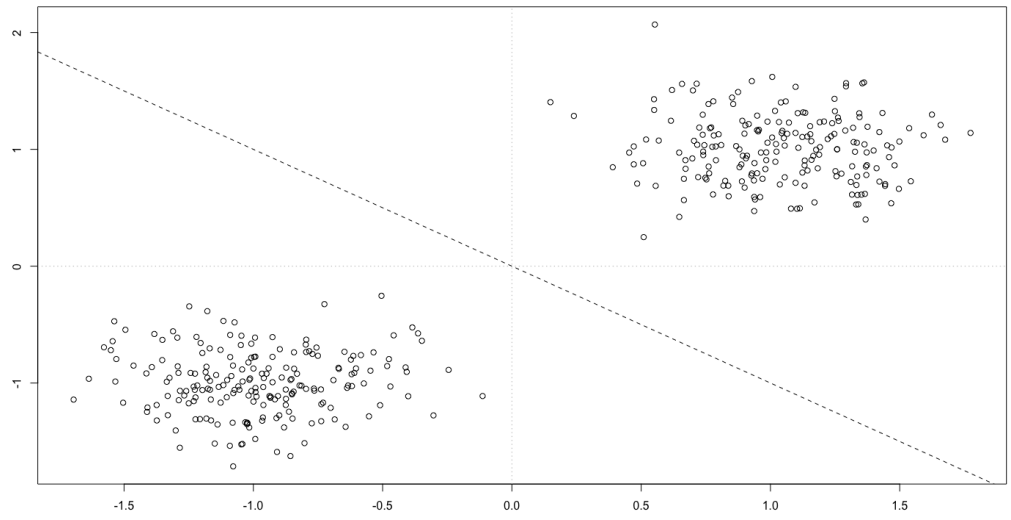


Figure 2: Randomly Generated Clustered Data
with Separating Hyperplane in $\mathbb{R}^2$

We see that there is a clear separation between the data points and a hyperplane that bisects the data. The equation of the hyperplane in this example is defined by $y = -x$. In the form $f(x) = \beta_0 + \beta \cdot x$, we have $f(x) = 0$, $\beta_0 = 0$, and $\beta = (1, 1)$ to give $0 = 0 + x_1 + x_2$ or $x_1 = -x_2$.

Notice that we can rotate the line slightly in either direction and still have separating hyperplane. So long as the line does not pass through the points that the gray line intersects, there exist infinitely many such separating hyperplanes. An example is shown in red in Figure **??**.

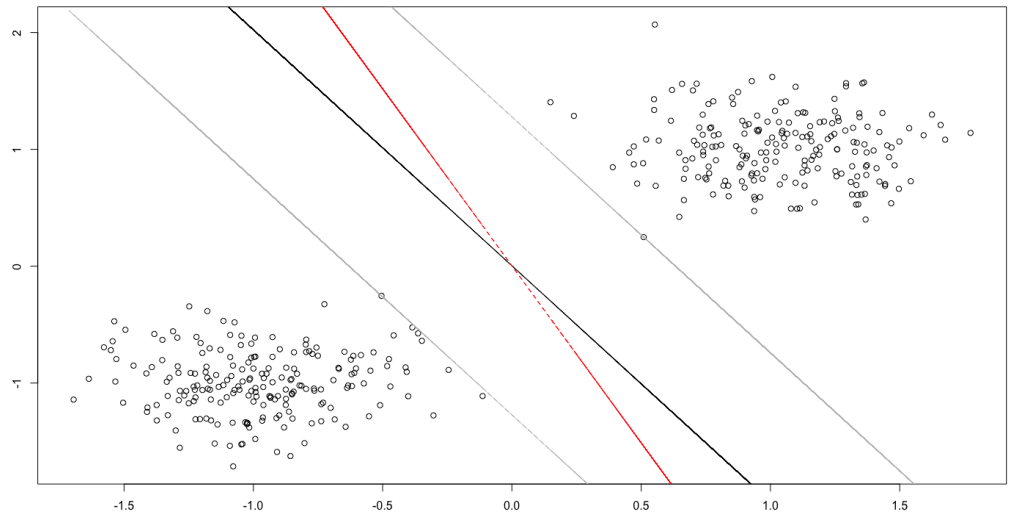


Figure 3: Randomly Generated Clustered Data with Multiple Hyperplanes in R^2

Similarly, in a three-dimensional space the hyperplane partitions the data. A plane separates the set into two distinct groups. An example can be generated with Python.

These functions set up our two data sets. The a, b, c values correspond to the x, y, z coordinates for each set. A randomly generated sample from the normal distribution is used. We then create the graph by setting up the axes and writing the equation for the hyperplane that will bisect the data before plotting the points.

The output is shown in Figure **??**.

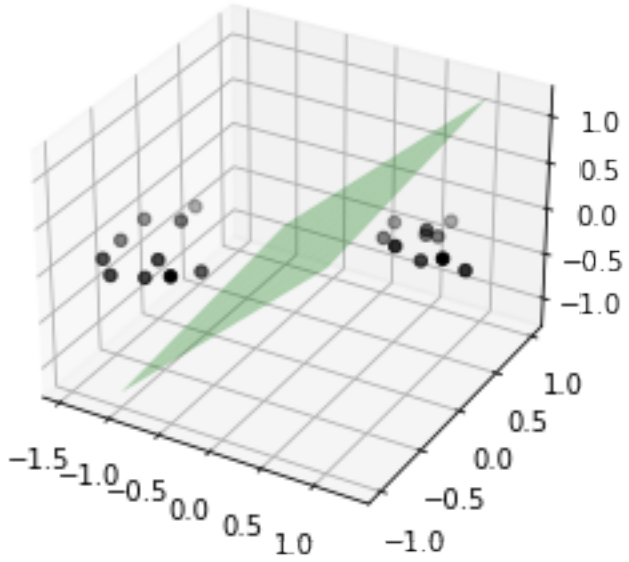


Figure 4: Randomly Generated Clustered Data
with Separating Hyperplane in $\mathbb{R}^3$

In this example, we see one set of points is to the left of the hyperplane while the other lies on the right. The equation of the hyperplane is $z = x + 0.15y$. In the form $f(x) = \beta_0 + \beta \cdot x$, we have $f(x) = 0$, $\beta_0 = 0$, and $\beta = (1, 0.15, -1)$ to give $0 = 0 + x_1 + 0.15x_2 - x_3$ or $x_3 = x_1 + 0.15x_2$. Like in the two-dimensional case, there is an infinite number of separating hyperplanes in three dimensions. By slightly rotating the plane shown in Figure **??** in either direction, we can visualize the concept of numerous other planes that would bisect the same data set.

3 MAXIMAL MARGIN CLASSIFIER

The use of hyperplanes to distinguish between two classes is central to the maximal margin classifier. Also called the optimal margin classifier, this method aims to find the separating hyperplane that maximizes the distance between the two sets of points. That is, if the plane were to rotate slightly in either direction, it would be closer to the set of data points. Our depiction of the maximal margin classifier draws upon those presented in **?** and **?**.

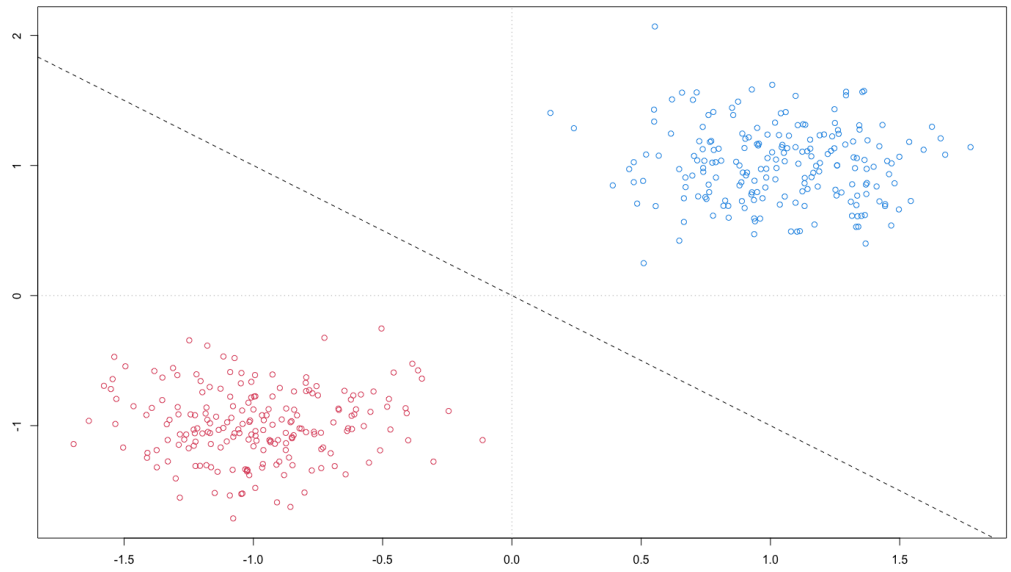


Figure 5: Randomly Generated Data with Associated Class Label Coloring

In the above example, we can classify each set of points to determine what side of the hyperplane they lie on. So, if $y_i = 1$ then $0 > \beta_0 + \beta \cdot x$ and if $y_i = -1$ then $0 < \beta_0 + \beta \cdot x$. Each y_i becomes the class label for that observation. Class labels are the identifiers for the data points and distinguish which class the observation belongs to. Introducing the following starred lines to our code produces colors in the graph according to the respective group:

The fifth line assigns a vector of 200 negative ones followed by 200 positive ones to the variable y. Combined with the adjusted plot function containing the color option $(3 - y)$, the first 200 points are plotted blue and the last 200 points are plotted red. This gives out two groups shown in Figure **??**. We need to find the closest blue and red points such that shifting the hyperplane in any such way results in a smaller margin for either group. These points are known as the support vectors, as they are the only points that end up formulating the hyperplane. At this point, we know that points with $0 > \beta_0 + \beta \cdot x$ will lie on one side of the hyperplane and have a class label y_i of 1, while the points $0 < \beta_0 + \beta \cdot x$ lie on the other and have a class label $y_i = -1$. Combining these two constraints, we can produce one inequality that

correctly classifies each point

$$y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) > 0$$

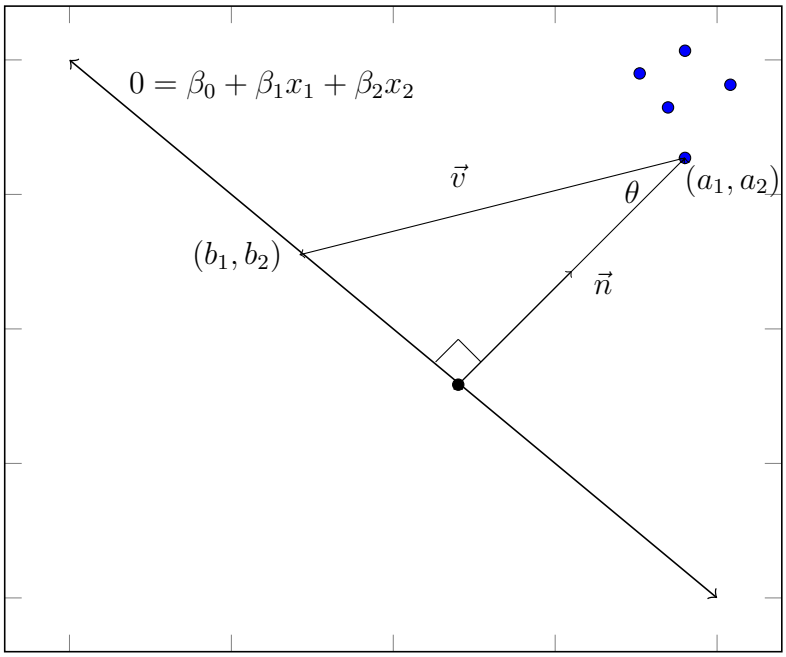


Figure 6: Diagram of Hyperplane and Support Vector

Suppose we look at a few points with a separating hyperplane of $0 = f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, pictured in Figure **??**. We assume the point (a_1, a_2) in the set of the positive class labels is the closest point to the set of points with a negative class label. Then, to compute the perpendicular distance d from the point (a_1, a_2) to $f(x_1, x_2) = 0$, we first find a normal vector to f. Through calculus, the gradient of the hyperplane calculated by (dx_1, dx_2) gives the normal vector $\vec{n} = \langle \beta_1, \beta_2 \rangle$ to f. Let (b_1, b_2) be any point on the hyperplane. Then $\vec{v} = (a_1 - b_1, a_2 - b_2)$. The distance

d equals $||\vec{v}|| \cos\theta$ through trigonometry. Then we have

$$\begin{aligned}
\vec{v} \cdot \vec{n} &= ||\vec{v}||||\vec{n}|| \cos\theta \\
\vec{v} \cdot \vec{n} &= d \cdot ||\vec{n}|| \\
d &= \frac{\vec{v} \cdot \vec{n}}{||\vec{n}||} \\
&= \frac{(a_1 - b_1, a_2 - b_2) \cdot \langle \beta_1, \beta_2 \rangle}{\sqrt{\beta_1^2 + \beta_2^2}} \\
&= \frac{\beta_1 a_1 - \beta_1 b_1 + \beta_2 a_2 - \beta_2 b_2}{\sqrt{\beta_1^2 + \beta_2^2}} \\
&= \frac{\beta_1 a_1 + \beta_2 a_2 - (\beta_1 b_1 + \beta_2 b_2)}{\sqrt{\beta_1^2 + \beta_2^2}}
\end{aligned}$$

Since (b_1, b_2) lies on the hyperplane, we have $\beta_0 + \beta_1 b_1 + \beta_2 b_2 = 0$ and $\beta_1 b_1 + \beta_2 b_2 = -\beta_0$. Thus,

$$\begin{aligned}
d &= \frac{\beta_1 a_1 + \beta_2 a_2 - (\beta_1 b_1 + \beta_2 b_2)}{\sqrt{\beta_1^2 + \beta_2^2}} \\
&= \frac{\beta_1 a_1 + \beta_2 a_2 - (-\beta_0)}{\sqrt{\beta_1^2 + \beta_2^2}} \\
&= \frac{\beta_1 a_1 + \beta_2 a_2 + \beta_0}{\sqrt{\beta_1^2 + \beta_2^2}}
\end{aligned}$$

Note that the numerator is exactly the equation f. In fact, there exist infinitely many such functions, as we could simply multiply f by a constant. Suppose we have the equivalent hyperplane $0 = C\beta_1 a_1 + C\beta_2 a_2 + C\beta_0$ for any $C \in \mathbb{R}$. Then,

$$\begin{aligned}
d &= \frac{C\beta_1 a_1 + C\beta_2 a_2 + C\beta_0}{\sqrt{(C\beta_1)^2 + (C\beta_2)^2}} \\
&= \frac{C(\beta_1 a_1 + \beta_2 a_2 + \beta_0)}{\sqrt{C^2(\beta_1^2 + \beta_2^2)}} \\
&= \frac{C(\beta_1 a_1 + \beta_2 a_2 + \beta_0)}{C\sqrt{\beta_1^2 + \beta_2^2}} \\
&= \frac{\beta_1 a_1 + \beta_2 a_2 + \beta_0}{\sqrt{\beta_1^2 + \beta_2^2}}
\end{aligned}$$

Notice that we are left with the same distance d, even with a different equation for the hyperplane. By imposing the condition $\beta_1^2 + \beta_2^2 = 1$, we restrict the above to just one possible hyperplane. So, we are left with

$$d = \frac{f(a_1, a_2)}{\sqrt{\beta_1^2 + \beta_2^2}}$$

We aim to find the greatest distance, or margin, by maximizing the coefficients to the function f of the closest point(s) to the hyperplane. Thus, we have the basic idea for the maximal margin classifier. We seek to maximize the margin M to find the best separating hyperplane given by

$$y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) \geq M$$

provided that

$$\sum_{j=1}^{p} \beta_j^2 = 1 \tag{1}$$

Though the maximal margin classifier is not a support vector machine quite yet, we may use the SVM module in the "e1071" library in R to determine the equation of the line.

Once downloading the package, we assign the x and y variables to one dataframe. We convert the y variable into a factor since class label is a binary value. In the **svm** function, the dataframe we just defined is the basis of the calculations. We assign the kernel as linear with a cost of 1000, though other options exist for non-separable data and non-linear decision boundaries which we will get to in later pages. We plot the function and see the output in Figure **??**.

As we see, the separating hyperplane is illustrated clearly albeit not pictured exactly linear. Details of the **svm** function can be found using the following:

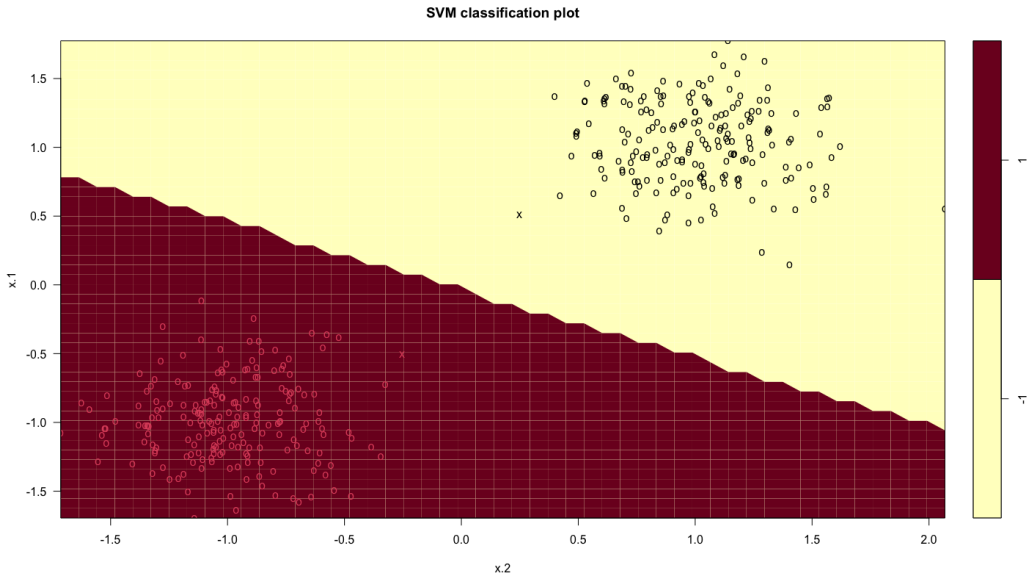


Figure 7: Maximal Margin SVM Module Output in R

```
> svmfit$rho
[1] 0.002291883
```

```
> svmfit$index
[1]  92 309
```

```
> svmfit$coefs
             [,1]
[1,]  1.561796
[2,] -1.561796
```

Figure 8: Selected Output of SVM Function

The **str** function tells us that **svmfit** is a list of 30 variables, some of which we defined in the function and others that were calculated. Particularly of interest are the **index**, **rho**, and **coefs** variables, which give the output in Figure **??**. The **rho** variable is the intercept of the separating hyperplane, and the **coefs** variable is an array of the coefficients of the supporting vectors. The **index** variable tells us the indices of the observations that the classifier depends on, or the support vectors. In this case, it is data points 92 and 309. If we were to rotate the hyperplane in any direction, the margin from these two points to the black line would inevitably become smaller. Visually, we can see these two points as those that the supporting hyperplanes pass through in Figure **??**.

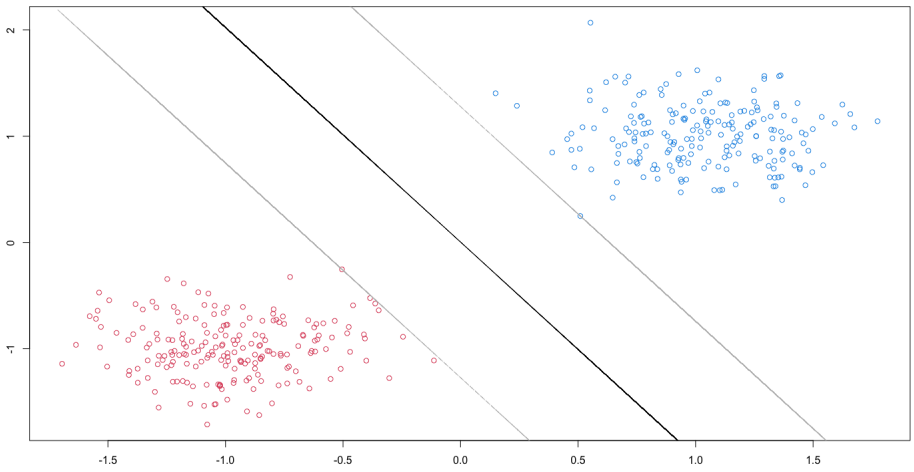


Figure 9: Maximal Margin Classifier Output in R

To find the equation of the line, we use the hyperplane function as outlined in **?**: In this function, **P** is the **svm** function which we labeled *svmfit*. The **data** is our dataframe containing the observations and class labels. **x** is the column vector of observations. We set **z**=0 for now, but this will later be shown to move the hyperplane vertically by a set constant.

Inside the function, the **alphas** are assigned to the coefficients of the hyperplane and multiplied by negative one. The **svs** variable takes the index of the support vectors and finds the exact point values in the data. **c** refers to the intercept and can be adjusted by the variable **z**. The variables **a** and **b** take the x and y coordinates, respectively, of the support vectors and multiply by their corresponding coefficient, as calculated by taking the transpose of the **alphas** variable, then summed. In Python, the **hyperplane** function looks slightly different due to the differences in the corresponding module, but the concept remains the same **?**.

The R code for the plot referenced in Figure **??** is below and pulls information from the **hyperplane** function.

4 SUPPORT VECTOR CLASSIFIER

What happens when data is not linearly separable? This occurs when the two groups of data have overlapping points. In this case, we cannot form a separating hyperplane as some points will be classified on the wrong side of the line. For non-linearly separable data, we introduce the support vector classifier. Much like the maximal margin classifier, we form a hyperplane between the groups of points in a data set. This classification method is also known as the soft-margin classifier since some points will not be classified on the correct side of the hyperplane unlike the maximal margin classifier. We will introduce the topic of costs, which essentially allow a certain number of observations to be classified incorrectly, then discuss the code in R and Python.

The following code in R can be used to plot non-linearly separable data with randomly generated points, shown in Figure **??**. We use the **set.seed()** function to replicate the randomly generated data. The process is the same as with linearly separable data, except we now narrow the gap between the means of the two groups which allows some overlap. Notice that we can no longer draw a straight line bisecting the two groups.

We can add a line to visualize a possible hyperplane by running the following line of code after we have constructed the graph.

As we can see, there is no possible linear hyperplane that separates the data. Some blue points are intermingled with the red and vice versa. To still construct a machine that attempts to classify the data, we introduce the topic of costs. In the maximal margin classifier, we aimed to find the greatest distance (margin) between the closest points of the two sets. Since there is no concrete margin in non-linearly separable data, we must introduce what is called a slack variable, denoted as ϵ_i for each observation x_i. For variables that are correctly classified, the corresponding observation's slack variable is 0. For points that are incorrectly classified, the slack variable takes on a value greater than 0 and incurs some cost for the maximization function. Since the slack variable is essentially a measure of how far a point lies on the wrong side of the margin, we constrict the sum of all slack variables to be less than a specified cost C. As we will see later, a user-defined cost can result in over- or under-fitting of the model. For a support vector classifier, our new maximization problem becomes

$$y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

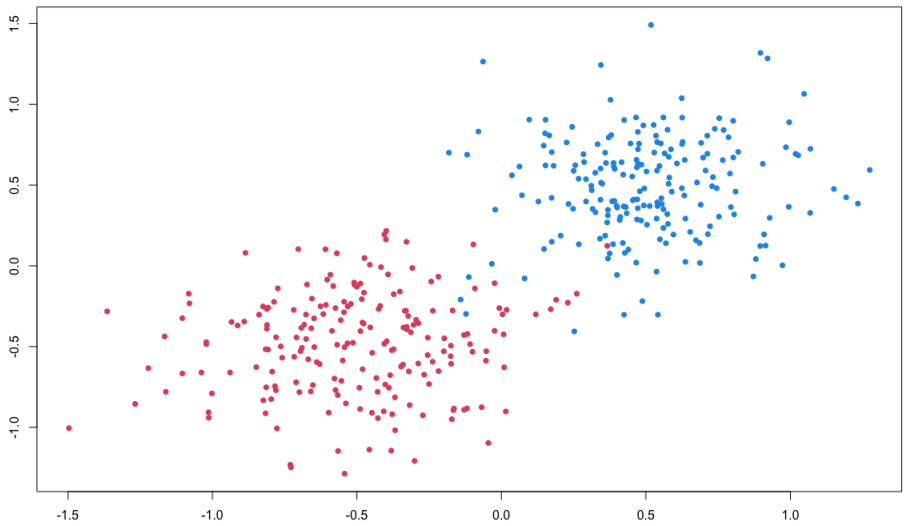


Figure 10: Randomly Generated Non-Linearly Separable Data

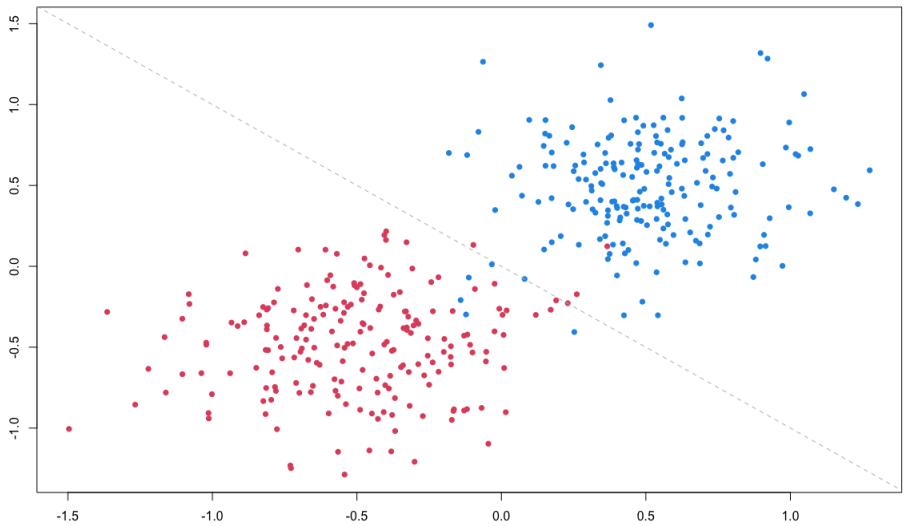


Figure 11: Randomly Generated Non-Linearly Separable Data With Line

provided that

$$\sum_{j=1}^{p} \beta_j^2 = 1 \tag{2}$$

$$\sum_{i=1}^{n} \epsilon_i \leq C, \epsilon_i \geq 0 \tag{3}$$

Using the same **svm** module as in the maximal margin classifier, we can find the best hyperplane to classify the data given a specified cost, in this case a value of 1000.

This produces the result in Figure **??**, graphically representing observations with slack variables greater than zero. Red points either inside the margin illustrated with the gray supporting hyperplanes or on the right side of the black hyperplane will have non-zero slack variables. Similarly, blue points inside the margin or on the left side of the hyperplane will have non-zero slack variables.

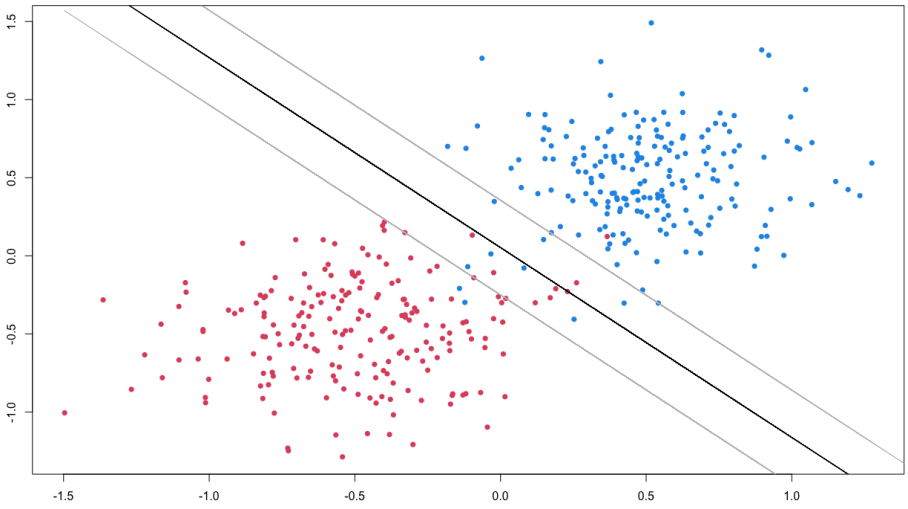


Figure 12: Randomly Generated Non-Linearly Separable Data With Line

Let us look at what happens with different values of the cost variable.

We see that as the cost increases, there are fewer support vectors. That is, fewer observations are classified incorrectly. Shouldn't it be the opposite? We see in **??** that the sum of the slack variables has to be less than or equal to the cost. A higher cost allows for greater slack variables, meaning more observations can lie outside the margin. This discrepancy lies in the underlying code. In both the R and Python modules for the support vector machine, the cost is inversely proportional to the C we defined in **??**.

The difference is a reference to the "dual problem", as there are two approaches to solving this problem. One approach, which we have previously outlined, fixes the variable C and the sum of β_j^2 in order to minimize the margin. We are left with solving

$$y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

subject to the conditions on β and ϵ. We could also solve this problem by switching our conditions and minimization as outlined in **?**. If we fix the margin to be 1, we aim to minimize $||\beta||$. The new problem becomes

$$\min ||\beta|| \text{ subject to } \begin{cases} y_i(x_i^T\beta + \beta_0) \geq 1 - \epsilon_i \forall i, \\ \epsilon_i \geq 0, \sum \epsilon_i \leq C, C \in \mathbb{R} \end{cases}$$

shouldn't that be opposite? larger cost means larger c so sum of ϵs can be greater so there can be more slack variables? look at page 389 for explanation

talk about how to test our model with prediction accuracy

overfitting vs underfitting in choosing cost

finding best cost through code function in e1071

code for python?

dual problem family of lines in the β space, all must lie on the right side of the line and we try to find the point closest to the origin

experiment with values of C and sum of slack variables; show on graph? trade-off between the two; $slack = 1 - y_i(f(x_i))$

minimize $||\beta^2|| + C\sum \epsilon_i$ subject to $y_i(\beta_i x_i + b) \geq 1 - \epsilon_i$ and $\epsilon_i \geq 0$

5 SUPPORT VECTOR MACHINE

- non-linear decision boundaries
- use of kernels
- multi-class data

6 EXAMPLE

- penguin/iris dataset: well-known example
- possible example with real implications

7 APPLICATIONS

One such example, as presented in *Intro to Statistical Learning with Applications in R* (**?**) is the use of support vector machines in genetic expression data. The *Khan* dataset in the $ISLR2$ package of R contains expression measurements for 2,308 genes from tissue samples of patients with one of four types of small round blue cell tumors. The data is then split up into training and test groups. There are 63 observations in the training set and 20 observations in the test set. It is not feasible to visually plot the data on a graph, as there is a very large number of features relative to the number of observations. However, due to the large number of features, it is easy to find a linearly separating hyperplane that predicts the type of cell tumor. As such, the SVM approach outlined in the textbook yields no data points that are misclassified in the training set and two test set errors. In this case, a support vector machine was used to classify and predict cancer types based on gene expression data.

In fact, SVMs are a popular choice in machine learning approaches to detecting cancer. Several studies have shown SVMs perform with great accuracy. Among those include the prediction of breast cancer using SVM and an extremely randomized trees classifier **?**

Support vector machines have also been used in multi-class lung cancer classification **?**.

8 ANALYSIS OF R AND PYTHON SVM MODULES AND DOCUMENTATION

- supporting documentation for each
- ease of use
- notes on function parameters