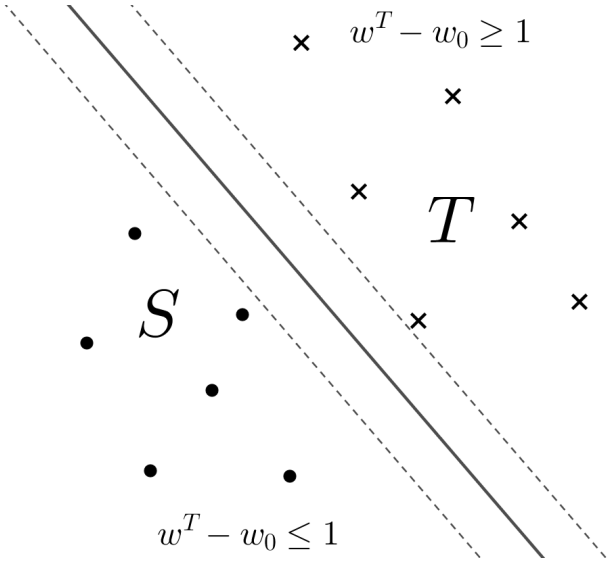
GEORGIA SOUTHERN UNIVERSITY

OPERATIONS RESEARCH



Classification Project

Samuel Walker
Abdullah Mamun
and Sattajit Sutradhar

December 4, 2017

Abstract

A surprising correlation between the levels of ribonucleic acid in three genes of a tick and the probability of the tick being a carrier of Lyme disease was discovered by researchers in the Department of Biology at Georgia Southern University. The objective of the researchers is to develop a classification procedure for determining the presence of Lyme disease in a tick by using the levels of ribonucleic in the three genes as an indicator. In standard classification problems a hyperplane is used to separate clusters, or sets, of data in a space of features. This paper will construct such a plane using a sample of forty ticks as a training set.

INTRODUCTION

Researchers are the Department of Biology discovered a surprising correlation between the levels of ribonucleic acid in three genes of tick and the probability said tick carries Lyme disease. A sample of forty ticks was examined and the various levels of ribonucleic acid for each indicator gene as well as the presence of Lyme disease in each tick was recored [1]. The data gathered will be used as training data for a supervised learning model which will assign ticks to a class based on whether or not they are carriers of Lyme disease.

This paper will first present a model for classifying the ticks by the presence of Lyme disease using a Support Vector Machine. For brevity, the mathematical derivation of this model will be relegated to Appendix A. Once our model has been properly defined we will present a solution to the initial classification problem using the training data. Afterwards we will analyze the solution using sensitivity analysis [**?**] to ensure that it is robust.

MODEL

Our objective is to determine a function which takes the levels of ribonucleic acid in the three indicator genes as input and outputs a binary hypothesis with 1 indicating the presence of Lyme disease and 0 otherwise. Define set S to be the set of all carriers of Lyme disease and set T to be all other ticks. Let $w, x \in \mathbb{R}^3$ and $w_0, b \in \mathbb{R}$. We want a hyperplane, $w^T x - w_0 = 0$, which partitions $\mathbb{R}^3$ into two cells: $w^T x - w_0 \geq b$ for $x \in S$ and $w^T x - w_0 \leq b$ for $x \in T$.

Perfect separation is not always guaranteed, some points belonging to either S or T will be a member of the wrong partition. In order to compensate for this error, we will introduce new variables y_i and z_i which measure the approximation error of our hypothesis for observation i. The positive constant b sets the width of the separation. Large values of b will catch more points which violate the inequality $w^T - w_0 \geq b$. For this paper we will set $b = 1$, other values of b will be considered in the section on analysis.

To find the best hyperplane, we look for the values of w and w_0 which minimize the approximation error. Thus we seek to solve

$$\begin{aligned} \text{Min } & \sum_{i=1}^{m} \frac{y_i}{m} + \sum_{i=1}^{k} \frac{z_i}{k} \\ \text{s.t. } & w^t s_i - w_0 + y_i \geq 1, \qquad \forall s_i \in S \\ & -(w^t t_i - w_0) + z_i \geq 1, \quad \forall t_i \in T \\ & y_i \geq 0,\ z_i \geq 0, \end{aligned}$$

SOLUTION

To solve this minimization problem we will implement the Simplex Algorithm using MATLAB. To do so we must re-write the problem in the standard format and use matrix notation. There are 20 observations in set S and 20 observations in set T. Thus let $\boldsymbol{I}$ be the identity matrix with 20 rows and columns, $\mathbf{0}$ be the zero matrix with 20 rows and columns, and $\mathbf{1}$ be a vector of all ones with 20 rows and 1 column. The model may then be written as

$$\begin{aligned} & \text{Min} \quad \sum_{i=1}^{m} \frac{y_i}{m} + \sum_{i=1}^{k} \frac{z_i}{k} \\ \text{s.t. } & \begin{bmatrix} -S & \mathbf{1} & -\boldsymbol{I} & \mathbf{0} \\ T & -\mathbf{1} & \mathbf{0} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} w \\ w_0 \\ y_i \\ z_i \end{bmatrix} \leq \begin{bmatrix} -\mathbf{1} \\ -\mathbf{1} \end{bmatrix} \\ & y_i \geq 0,\ z_i \geq 0. \end{aligned}$$

[1] this data can be found in Appendix B

Running MATLAB's linprog method, which uses the 'dual-simplex' algorithm, we found that w^* and w_0^*, the optimal values of w and w_0, were $w* = \langle -19.6344, -6.6928, -3.9217 \rangle$ and $w_0^* = -160.4726$.

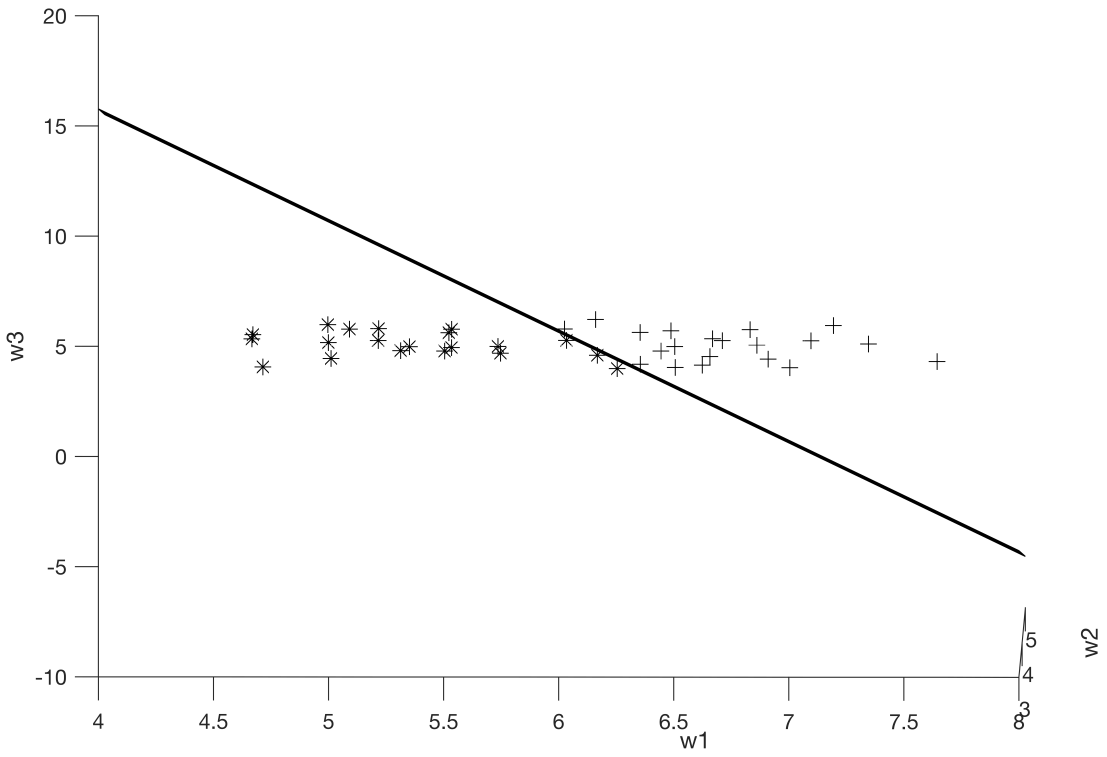


Figure 1: Ticks which are known carriers of Lyme disease are shown with stars. The hyperplane perfectly partitions the two sets.

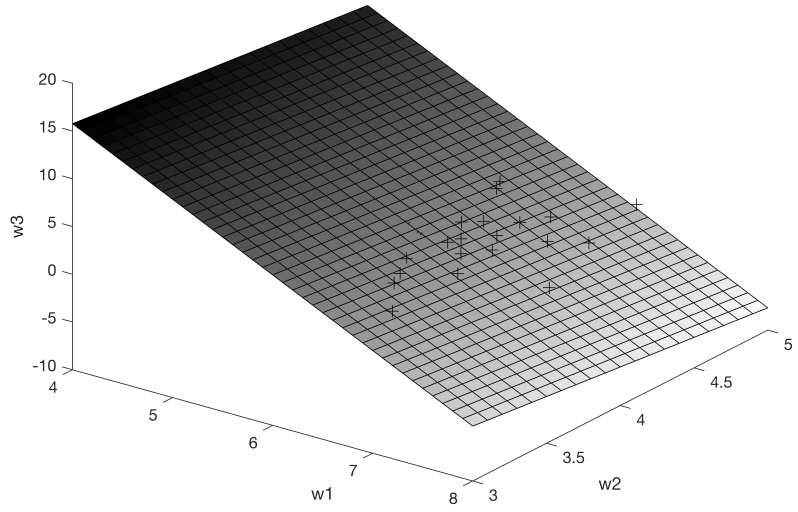


Figure 2: Another view of the hyperplane with ticks who are not carriers being displayed.

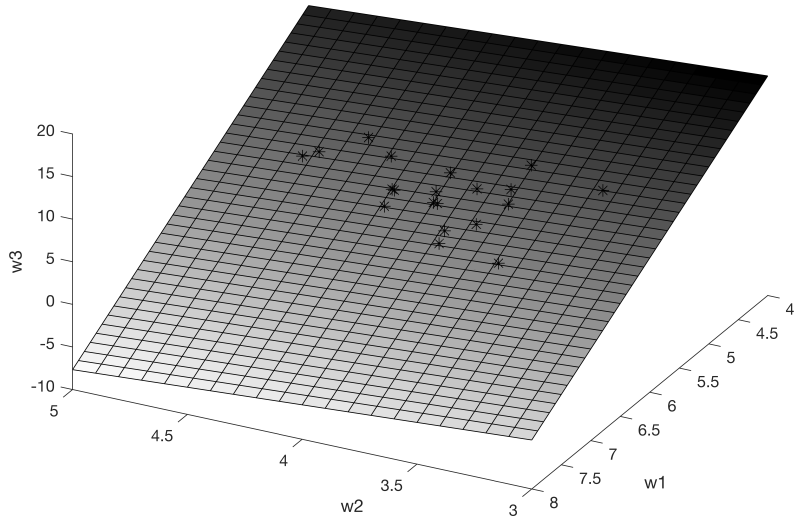


Figure 3: Opposite view to the previous figure. Ticks displayed here are known carriers of Lyme disease.

ANALYSIS

The shadow prices for the constraint equations were all zero. This indicates that scaling b has no affect on the slope on the hyperplane as is expected. Scaling b will only scale the values of w and w_0.

The values of w and w_0 are highly sensitive. MATLAB gives an allowable decreases/increase of zero for each of $w_0, \ldots, w_3$. This is likely the product of MATLAB's machine precision not being high enough to quantify small perturbations of w or w_0. This result makes sense because the separation of the two sets is weak.

CONCLUSIONS

Using linear programming we were able to develop a classification plane which partitions the feature space into two bins. For every observation our plane produces a hypothesis on which bin the observation belongs - if the observation is a carrier of Lyme disease then it is placed in the first bin. Two two sets of observations are weakly separated. That is, points from both sets lie close to the separating hyperplane. The data in the sample collected was linearly separable; Mercer kernels were not required to find a feasible linear hypothesis.

The shadow prices for each of the forty constraints were zero. Any change in b, the width of the offender net around the hyperplane, does not affect the end hypothesis. This makes sense since any change in the values of b can be reflected by arbitrarily scaling the values of w and w_0. This scaling will not affect the slope of the plane. Sensitivity analysis revealed that the plane is highly sensitive to changes in w and w_0. Any perturbations in the values of $w_0, \ldots, w_3$ has a significant impact in the hypothesis of our classification hyperplane. Again, this is due to the weak separation of the two data sets.

BIBLIOGRAPHY

[1] Hillier, Frederick S. *Introduction to Operations Research*. McGraw-Hill. 1989.

1

Appendices

APPENDIX A

To derive the mathematical formulation of our minimization problem we will first define a classification problem in Euclidean space.

Theorem 1. *Let $S, T \subseteq \mathbb{R}^n$ be finite sets, say $|S| =$ and $|T| = k$. The classification problem consists in finding a function $f : \mathbb{R}^n \to \{-1, 1\}$ such that*

$$f(x) = \begin{cases} 1, & \text{if } x \in S \\ -1, & \text{if } x \in T. \end{cases}$$

A possible solution to this problem is to define f as a hyperplane. Thus our problem is to find a hyperplane $w^T x - w_0 = 0$, where $w \in \mathbb{R}^n$ is a constant vector with n components, $w_0 \in \mathbb{R}$, and $x, b \in \mathbb{R}^n$ is a variable with n components, such that

$$\begin{aligned} w^T s - w_0 &\geq b, \quad \forall s \in S \\ w^T t - w_0 &\leq -b, \quad \forall t \in T. \end{aligned}$$

In our formulation we set $b = 1$. In standard form:

$$\begin{aligned} -w^T s + w_0 &\leq -b, \quad \forall s \in S \\ w^T t - w_0 &\leq -b, \quad \forall t \in T. \end{aligned}$$

Some points will not be classified correctly. We call such points "offenders". For each points in S and T we will associate a variable $y_i \geq 0$ and $z_i \geq 0$ such that

$$\begin{aligned} -w^T s + w_0 - y_i &\leq -b, \quad \forall s \in S \\ w^T t - w_0 - z_i &\leq -b, \quad \forall t \in T. \end{aligned}$$

These values correspond to the error in classification by our hypothesis. We wish to minimize the average error. Hence our problem is:

$$\begin{aligned} \text{Min } & \sum_{i=1}^{m} \frac{y_i}{m} + \sum_{i=1}^{k} \frac{z_i}{k} \\ \text{s.t. } & w^t s_i - w_0 + y_i \geq 1, && \forall s_i \in S \\ & -(w^t t_i - w_0) + z_i \geq 1, && \forall t_i \in T \\ & y_i \geq 0,\ z_i \geq 0. \end{aligned}$$

APPENDIX B

Table 1.1: S

x_1	x_2	x_3	ID
5.202898	3.956818	3.758593	1
5.067827	4.67962	3.137106	1
4.983422	4.129856	3.396585	1
6.156168	3.877599	3.224985	1
5.297728	4.111029	3.057959	1
5.7357	3.823527	3.395893	1
6.246019	3.595627	3.064928	1
4.707389	3.536235	3.225886	1
4.649478	4.571746	3.054629	1
5.505461	4.239719	3.670282	1
5.520044	4.048321	3.297742	1
5.19339	4.71956	3.100722	1
5.718818	4.227902	3.071904	1
6.021652	3.888151	3.880441	1
5.341147	3.785079	3.754247	1
5.517542	4.246328	3.816606	1
4.654915	3.861604	3.98214	1
5.489728	4.073375	3.1036	1
4.998947	3.861171	3.099042	1
4.97783	4.388809	3.798832	1

Table 1.2: T

x_1	x_2	x_3	ID
6.651548	4.233757	3.414735	0
7.633764	3.767147	3.119405	0
6.618614	3.404167	3.52603	0
6.499548	3.52558	3.225073	0
7.080586	4.131223	3.486412	0
6.003723	4.538547	3.370215	0
6.33943	4.050519	3.98335	0
6.432171	3.893466	3.388319	0
6.645498	3.838292	3.229245	0
6.84857	3.912995	3.623298	0
6.345207	3.673677	3.13654	0
6.47219	4.109972	3.967469	0
6.49126	3.943486	3.515072	0
7.329364	4.242249	3.16307	0
6.696751	4.047553	3.621902	0
6.14046	4.422031	3.985954	0
6.813471	4.333092	3.668772	0
6.901336	3.644052	3.418916	0
7.170416	4.672808	3.323345	0
7.002646	3.132415	3.835255	0