Predictive Analytics for Bankruptcy: Utilizing SVM and PGD to Forecast Financial Distress in Companies

Huang Tsz Wing

10th May 2024

1 Background and Introduction

In this project, we are going to develop a bankruptcy prediction models to identify early warning signals of financial distress to enable proactive decision-making and risk mitigation. The goal of our project was to develop a classifier capable of predicting whether a company will go bankrupt based on its recent performance, utilizing data from approximately 10,000 Polish companies spanning from 2000 to 2012. To achieve this objective, we will use a powerful supervised algorithm ,Support Vector Machines (SVM), which is suitable for this task due to its effectiveness in handling high-dimensional data and their capability to model non-linear decision boundaries. Moreover, we will use Projected Gradient Decent (PGD) to optimized the performance of the SVM. In this project, the dataset has 64 Attributes/Features (Appendix A) and We will make use of the Julia1 environment.

2 Model and Theory

2.1

• The following are given. Let $d \in \mathbb{N}$ be the dimension of the feature vector, the (linear) classifier is described by the tuple (w, b) where $w \in \mathbb{R}^d$ is the direction parameters and $d\mathbb{R}$ is the bias parameter, both specifying the linear model. Given (w, b), a feature vector $x \in \mathbb{R}^d$ is classified into a label $y\{\pm 1\}$ such that

$$y = \begin{cases} +1, if \quad w^\top x + b = w_1x_1 + ... + w_dx_d + b \geq 0, \\ -1, if \quad w^\top x + b = w_1x_1 + ... + w_dx_d + b < 0 \end{cases} \tag{1}$$

Denote the training dataset with m samples as $\{x^{(i)}, y^{(i)}\}_{i=1}^m$, where $x^{(i)} \in \mathbb{R}^d$ is thje ith feature vector with d attributes and $y^{(i)} \in \{\pm 1\}$ is the associated lable. Let $R_0 > 0$ and $\ell_i > 0, i = 1, ..., m$ be a set of positive weights. The following optimization problem designs a soft-margin classifier:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \sum_{i=1}^m \ell_i \max\{0, 1 - y^{(i)}((x^{(i)})^\top w + b)\} \quad s.t. \quad w^\top w \leq R_0 \qquad (2)$$

In (1), the term $\max\{0, 1 - y^{(i)}((x^{(i)})^\top w + b)\}$ evaluates the amount of error for ith sample. The point is mis-classified if and only if $y^{(i)}((x^{(i)})^\top w + b) < 1$, then the term is > 0 . If the optimal objective value of (2) is zero, that means in the summation of the objective function, all terms are 0. Which indicate that there are zero error among the data points. Therefore, any optimal solution to (2) is a classifier (w^*, b^*) that can correctly distinguish the m training samples into the +1 or -1 lables.

• The following is a example of training dataset with $d = 2$ where the optimal value of (2) is not zero. In Figure 1, point O and point N are mis-classified. For point O, $y(O) = 1, max\{0, y^{(O)}((x^{(O)})^\top w + b)\} > 0$. For point N, $y(N) = -1, max\{0, y^{(N)}((x^{(N)})^\top w + b)\} > 0$. Therefore the objective value is > 0, i.e. not equal to zero.

The following steps are rewriting the optimization problem in to an equivalent nonlinear program, e.g., $\min f_0(x)\ s.t.\ f_i(x) \leq 0,\ i = 1, ...m$. This process we take out the max function, which make the function differentiable.

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \sum_{i=1}^m \ell_i \max\{0, 1 - y^{(i)}((x^{(i)})^\top w + b)\} \quad s.t. \quad w^\top w \leq R_0$$

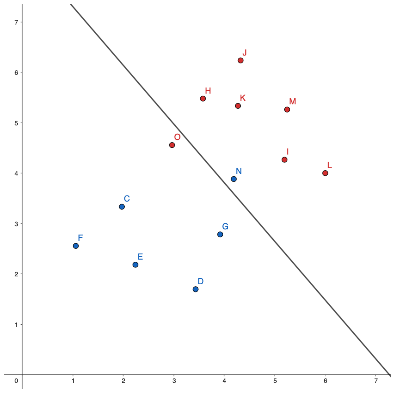

Figure 1: Example of dataset with optimal value $\neq 0$

$$
\begin{gathered}
\leftrightarrow \quad \min_{w \in \mathbb{R}^d, b \in \mathbb{R}, t \in \mathbb{R}, h \in \mathbb{R}} t \quad s.t. \quad t \geq \sum_{i=1}^{m} \ell_i h_i, \\
h_i \geq \max\{0, 1 - y^{(i)}((x^{(i)})^\top w + b),\ i = 1, ..., m \\
w^\top w \leq R_0 \\
\leftrightarrow \quad \min_{w \in \mathbb{R}^d, b \in \mathbb{R}, t \in \mathbb{R}, h \in \mathbb{R}} t \quad s.t. \quad t \geq \sum_{i=1}^{m} \ell_i h_i, \\
h_i \geq (1 - y^{(i)}((x^{(i)})^\top w + b)),\ i = 1, ..., m \\
h_i \geq 0,\ i = 1, ..., m, \quad w^\top w \leq R_0 \\
\leftrightarrow \quad \min_{w \in \mathbb{R}^d, b \in \mathbb{R}, t \in \mathbb{R}, h \in \mathbb{R}} t \quad s.t. \quad \sum_{i=1}^{m} \ell_i h_i - t \leq 0, \\
(1 - y^{(i)}((x^{(i)})^\top w + b)) - h_i \leq 0,\ i = 1, ..., m \\
-h_i \leq 0,\ i = 1, ..., m, \quad w^\top w - R_0 \leq 0
\end{gathered} \tag{3}
$$

• The following process aims to derive the Karush-Kuhn-Tucker (KKT) Conditions for the equivalent formulation in (3)

Lagrangian function:

$$L(t, z^{(i)}, x^{(i)}, y^{(i)}, \mu_1, \mu_{2,1}, ..., \mu_{2,m}, \mu_{3,1}, ..., \mu_{3,m}, \mu_4) =$$

$$t + \mu_1(\sum_{i=1}^{m} \ell_i h_i - t) + \mu_{2,1}(1 - y^{(i)}((x^{(i)})^\top w + b) - h_i) + ... + \mu_{2,m}(1 - y^{(i)}((x^{(i)})^\top w + b) - h_i)$$

$$+\mu_{3,1}(-h_i) + ... + \mu_{3,m}(-h_i) + \mu_4(w^\top w - R_0)$$

KKT Condition:

1. First-Order Necessary Conditions (FONC)

$$\frac{\partial L}{\partial w} = -\mu_{2,1} y^{(i)} \sum_{j=1}^{d} x_j^{(i)} - ... - \mu_{2,m} y^{(i)} \sum_{j=1}^{d} x_j^{(i)} + 2\mu_4 \sum_{j=1}^{d} w_j = 0$$

$$\frac{\partial L}{\partial b} = -\mu_{2,1}y^{(i)} - ... - \mu_{2,m}y^{(i)} = 0$$

$$\frac{\partial L}{\partial t} = 1 - \mu_1 = 0$$

$$\frac{\partial L}{\partial h} = \mu_1 \sum_{i=1}^{m} \ell_i - \mu_{2,1} - ... - \mu_{2,m} - \mu_{3,1} - ... - \mu_{3,m} = 0$$

2. Slackness

$$\mu_1(\sum_{i=1}^{m} \ell_i h_i - t) = 0$$

$$\mu_{2,1}(1 - y^{(i)}((x^{(i)})^\top w + b) - h_i) = 0$$

$$...$$

$$\mu_{2,m}(1 - y^{(i)}((x^{(i)})^\top w + b) - h_i) = 0$$

$$\mu_{3,1}(-h_i) = 0$$

$$...$$

$$\mu_{3,m}(-h_i) = 0$$

$$\mu_4(w^\top w - R_0) = 0$$

3. Primal constraint

$$\sum_{i=1}^{m} \ell_i h_i - t \leq 0$$

$$(1 - y^{(i)}((x^{(i)})^\top w + b)) - h_i \leq 0,\ i = 1, ..., m$$

$$-h_i \leq 0,\ i = 1, ..., m,$$

$$w^\top w - R_0 \leq 0$$

4. Dual constraint

$$\mu_1, \mu_{2,1}, ..., \mu_{2,m}, \mu_{3,1}, ..., \mu_{3,m}, \mu_4 \geq 0$$

- Suppose the optimal value of (2) is zero, the following show that there may exist more than one optimal solution to (2). Let $d = 2$ and consider Figure 2. There are four classifier which correctly classified all the data points. Since there are no error, the value of (2) are zero. With this example, we can see that there may exist more than one optimal solution to (2) with the optimal value 0.

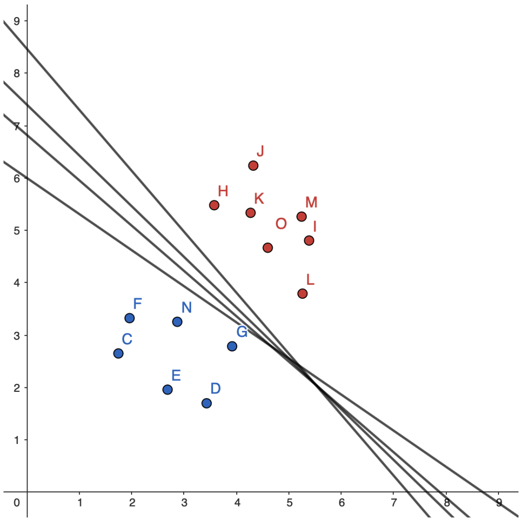

Figure 2: Example with optimal value = 0

The following explain why there may exist more than one optimal solution to (2) using the KKT conditions derived in the above.

From $w^\top w - R_0 \leq 0$ and $\mu_4(w^\top w - R_0) = 0$
If $\mu_4 > 0$, then $w^\top w = R_0 \Leftrightarrow \sum_{i=1}^{d} w_i^2 = R_0$, so that w can be positive or negative, which may have different solution
If $\mu_4 = 0$, then $w^\top w <= R_0$, which means that allow any w with a norm up to $\sqrt{R_0}$. From this observation, we can see that w may have different solution as long as it satisfy all others condition and $\leq R_0$. Therefore, there may exist more than one optimal solution to (2).

2.2

From the above process, the optimization problem (2) is rewritten into nonlinear problem shown in (3). The following process is to rewrite as a Second-order Cone Programming (SOCP) problem:

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, t\in\mathbb{R}, h\in\mathbb{R}} t \quad s.t. \quad \sum_{i=1}^{m} \ell_i h_i - t \leq 0, \quad (1 - y^{(i)}((x^{(i)})^\top w + b)) - h_i \leq 0,\ i = 1, ..., m$$

$$-h_i \leq 0,\ i = 1, ..., m, \quad w^\top w - R_0 \leq 0$$

Consider each constraints:
(1): $\sum_{i=1}^{m} \ell_i h_i - t \leq 0$
Since ℓ_i is a constant, h_i and t is a vector, this is a linear constraint.

(2): $(1 - y^{(i)}((x^{(i)})^\top w + b)) - h_i \leq 0,\ i = 1, ..., m$
$\Leftrightarrow 1 - y^{(i)} x^{(i)\top} w + y^{(i)} b - h_i \leq 0, i = 1, ..., m$
Since $y^{(i)}$ and $x^{(i)}$ are constants, $b \in \mathbb{R}$ is a bias parameter, and $h_i \in \mathbb{R}$ is a variable. Clearly, this is a linear constraint.

(3): $-h_i = 0$
Obviously, this is a linear constraint.

(4): $w^\top w - R_0 \leq 0$
$\Leftrightarrow \sum_{j=0}^{d} w_j^2 \leq R_0$
$\Leftrightarrow \sqrt{w_1^2 + ... + w_d^2} \leq \sqrt{R_0}$
$\Leftrightarrow ||w|| \leq \sqrt{R_0}$
With the above observation, the problem is now a SOCP problem.

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, t\in\mathbb{R}, h\in\mathbb{R}} t \quad s.t. \quad \sum_{i=1}^{m} \ell_i h_i - t \leq 0,$$
$$(1 - y^{(i)}((x^{(i)})^\top w + b)) - h_i \leq 0,\ i = 1, ..., m$$
$$-h_i \leq 0,\ i = 1, ..., m, \quad ||w|| \leq \sqrt{R_0} \tag{4}$$

• In the following process, the shaping constraint will be incorporate into the soft-margin problem.
The directional parameters$w \in R^d$ satisfies the following shaping constraint:

$$w^\top \Sigma w + c^\top w \leq R_0$$

where $\Sigma \in \mathbb{R}^{d\times d}$ is a given symmetric, positive definite matrix, and $c \in \mathbb{R}^d$ is a given vector.

The directional parameter and bias parameter belongs to an ℓ_1 ball to promote sparsity, i.e.,

$$\sum_{j=1}^{d} |w_j| + |b| \leq R_1$$

Now we can formulate the problem:

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, t\in\mathbb{R}, h\in\mathbb{R}} t \quad s.t. \quad \sum_{i=1}^{m} \ell_i h_i - t \leq 0,$$
$$(1 - y^{(i)}((x^{(i)})^\top w + b)) - h_i \leq 0, \; i = 1, ..., m$$
$$-h_i \leq 0, \; i = 1, ..., m, \quad ||w|| \leq \sqrt{R_0}$$
$$w^\top \Sigma w + c^\top w \leq R_0$$
$$\sum_{j=1}^{d} |w_j| + |b| \leq R_1$$

The following process are to rewrite the problem into SOCP. Consider the two new constraint:

(5) $w^\top \Sigma w + c^\top w \leq R_0$
$\Leftrightarrow$ Let $k \geq w^\top \Sigma w$, and $k + c^\top w \leq R_0$
$\Leftrightarrow w^\top \Sigma w - k \leq 0$
$\Leftrightarrow 4w^\top \Sigma w - 4k \leq 0$
$\Leftrightarrow 4w^\top \Sigma w + (1-k)^2 - (1+k)^2 \leq 0$
$\Leftrightarrow 4w^\top \Sigma w + (1-k)^2 \leq (1+k)^2, \quad 1+k \geq 0$
$\Leftrightarrow \sqrt{4w^\top \Sigma w + (1-k)^2} \leq (1+k), \quad 1+k \geq 0$
$\Sigma = (\Sigma^{\frac{1}{2}})^\top \Sigma^{\frac{1}{2}}$
$\Leftrightarrow \sqrt{4w^\top (\Sigma^{\frac{1}{2}})^\top \Sigma^{\frac{1}{2}} w + (1-k)^2} \leq (1+k), \quad 1+k \geq 0$
$\Leftrightarrow \left\| \begin{bmatrix} 2\Sigma^{\frac{1}{2}} w \\ 1-k \end{bmatrix} \right\| \leq 1+k, \quad 1+k \geq 0$

(6) $\sum_{j=1}^{d} |w_j| + |b| \leq R_1$
Let $-w_j \leq u_j \leq w_j, \quad j = 1, ..., d, \quad$ and $-v \leq b \leq v$
$\Leftrightarrow \sum_{j=1}^{d} u_j + v \leq R_1$
$-w_j \leq u_j, \quad u_j \leq w_j, \; j = 1, ..., d$
$-v \leq b, \quad b \leq v$

Now we have rewrite the problem in to SOCP:

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, t\in\mathbb{R}, h\in\mathbb{R}, k\in\mathbb{R}, u\in\mathbb{R}^d, v\in\mathbb{R}} \quad t$$
$$s.t. \quad \sum_{i=1}^{m} \ell_i h_i - t \leq 0$$
$$(1 - y^{(i)}((x^{(i)})^\top w + b)) - h_i \leq 0, \; i = 1, ..., m$$
$$-h_i \leq 0, \; i = 1, ..., m, \quad ||w|| \leq \sqrt{R_0}$$
$$k + c^\top w \leq R_0, \quad \left\| \begin{bmatrix} 2\Sigma^{\frac{1}{2}} w \\ 1-k \end{bmatrix} \right\| \leq 1+k, \quad 1+k \geq 0$$
$$\sum_{j=1}^{d} u_j + v \leq R_1, \quad -w_j \leq u_j, \quad u_j \leq w_j, \; j = 1, ..., d$$
$$v \geq b, \; v \geq -b \tag{5}$$

In the following process, a Mixed Integer Program (MIP) will be formulated, which imposes a hard constraint on the sparsity of the classifier, i.e., for given $R_0 > 0, R_1 > 0, S > 0$.
The directional parameters $w \in \mathbb{R}^d$ satisfies the following shaping constraint:

$$w^\top \Sigma w + c^\top w \leq R_0$$

where $\Sigma \in \mathbb{R}^{d\times d}$ is a given symmetric, positive definite matrix, and $c \in \mathbb{R}^d$ is a given vector. Each element in w is bounded such that

$$-R_1 \leq w_j \leq R_1,\ j = 1, ..., d$$

The number of non-zero elements in the vector w is constrained such that

$$(\text{no. of non-zero elements in the vector } w) \leq S$$

An integer variable $z_j \in \{0, 1\},\ j = 1, ..., d$ are introduce here.

$$z_j = \begin{cases} 0, if & w_j = 0 \\ 1, if & w_j \neq 0 \end{cases}$$

Then the problem can be reformulate into :

$$\begin{gathered}
\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, t\in\mathbb{R}, h\in\mathbb{R}} t \quad s.t. \quad \sum_{i=1}^{m} \ell_i h_i - t \leq 0, \\
(1 - y^{(i)}((x^{(i)})^\top w + b)) - h_i \leq 0,\ i = 1, ..., m \\
-h_i \leq 0,\ i = 1, ..., m, \quad ||w|| \leq \sqrt{R_0} \\
w^\top \Sigma w + c^\top w \leq R_0 \\
-R_1 z_j \leq w_j \leq R_1 z_j,\ j = 1, ..., d \\
\sum_{j=1}^{d} z_j \leq S \\
R_0, R_1, S > 0, \quad z_j \in \{0, 1\},\ j = 1, ..., d
\end{gathered} \tag{6}$$

3 Experiments

3.1 In this section, the above optimization designs will be put into practice. In the following, dateset of $m = 20$ companies will be focused. Each company has 64 attributes (performance indicators). The dataset also contains information of whether the company has bankrupted or not, treated as the label $y_i \in \{\pm 1\}$. If the company is bankrupt, the datapoint is mark in red, otherwise, the datapoint is mark in blue.
First, choosing Attr1 (net profit / total assets) and Attr2 (total liabilities / total assets):

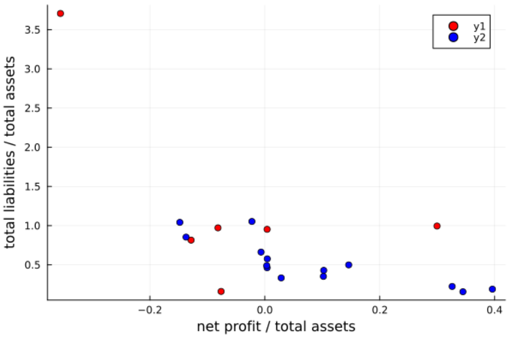

Figure 3: Scatter plots with Attr1 & Attr2

From Figure 3, the data points are scattered across the plot without a clear linear pattern. Most of the blue points (y2) are concentrated towards the center-bottom section of the plot, with x-values ranging roughly from -0.2 to 0.2 and y-values mainly between 0 and 1. The red points (y1) are fewer and more spread out,the highest red points reach y-values up to around 3.5. Therefore, there is no apparent trend or correlation that can be observed from the scatter plot.

The following plots are using different attributes:

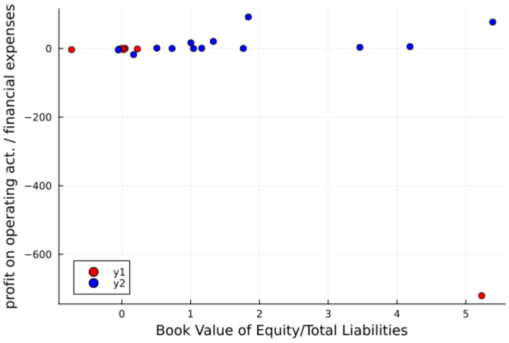

Figure 4: Scatter plots with Attr8 & Attr27

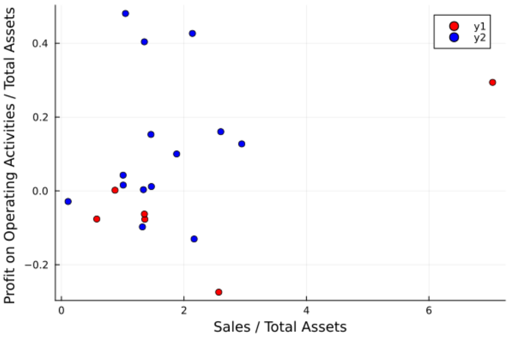

Figure 5: Scatter plots with Attr9 & Attr22

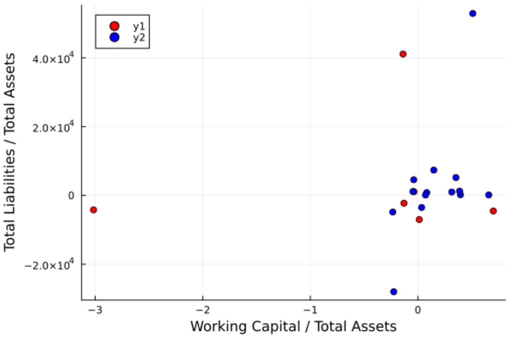

Figure 6: Scatter plots with Attr3 & Attr15

The above three scatter plots present relationships between various financial ratios: the first plot correlates 'Book Value of Equity/Total Liabilities' with 'profit on operating activities/financial expenses', the second correlates 'Sales / Total Assets' with 'Profit on Operating Activities / Total Assets', and the third correlates 'Working Capital / Total Assets' with 'Total Liabilities / Total Assets'. Across these plots, although Figure 4 seems to can have a linear classifier, but it still hard to perfectly classify. In Figure 5 and Figure 6, y1 and y2 are even more spread out. From these observation, there is no apparent trend or correlation that can be observed from the scatter plots. So a linear SVM classifier would struggle to perfectly classify the groups without mis-classification, as the data does not support a simple linear division.

3.2

• In the following, we are going to Implement and solve the optimization problem (4), (5), and (6) using Julia with the solver ECOS and JuMP for SOCP, and ECOS and JuMP for MIP. As we can see in Section 3.1, using 2 attributes is difficult to classify two groups of data. Therefore, we choose attributes 1-10 for this time.

First, in the program, we convert all the variables, constraints and objective function into the program. You may refer to the program to look for the details. After solving (3),(4), and (6), we now have the optimal solution(weights = 1). The following graph reveal the value of classifier $i.e.w_1, ..., w_{10}$.

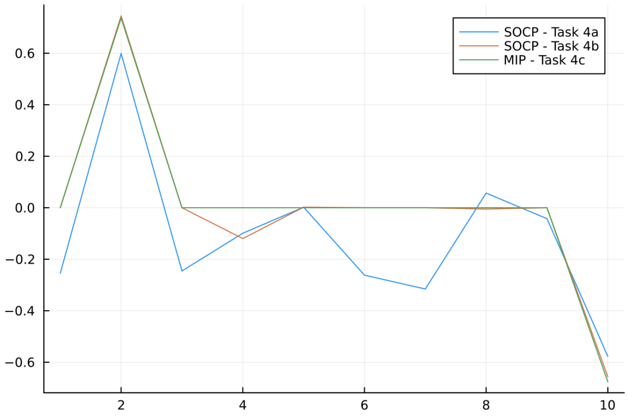

Figure 7: Classifier value among 3 program

From Figure 7, we can observe that 3 problem also show a peak at the 2^{nd} feature, and a bottom at the 10^{th} feature indicating that this feature has a significant weight in the classification decision for these problem. Among 3 problem, we can see that the classifier for task 4a is the least sparse, which indicate that most features have some level of influence on the model. The classifier for task 4c is the sparsest, which indicate only a few features in making its classification decisions. So we can see that, $Attr2$ (total liabilities / total assets) and $Attr10$ (equity/total assets) is the most significant features among these program, which is reasonable. $Attr2$ is the ratio that measure the leverage. A high leverage ratio imply that the company is heavily reliant on debt financing, which may increase the risk of bankruptcy if the company cannot meet its liabilities. $Attr10$ is the ratio indicates the proportion of a company's assets are financed by shareholders' equity. A low ratio indicate more assets are financed by liabilities than equity. These two attributes indicate the equity and liabilities ratio of a company, which are reasonable features to predict a company is bankrupted or not.

3.3

• Next, we would like to evaluate the error performance of the classifier with different weights. This can be evaluated by the error rate when the classifier applied on a certain set of data. It can further be specified into false alarm rate and missed detection rate. To describe these metrics, note for a given classifier (w, b), the predicted label is

$$\hat{y}^{(i)} = \begin{cases} +1, if & w^\top x^{(i)} + b \geq 0, \\ -1, if & w^\top x^{(i)} + b < 0 \end{cases}$$

With the training dataset $\{x^{(i)}, y^{(i)}\}_{i=1}^m$. Suppose that m_- is the number of samples with $y_i = -1$ and $\mathbb{D}_-$ is the corresponding set of samples, m_+ is the number of samples with $y_i = 1$ and $\mathbb{D}_+$is the corresponding set of samples. The error rates are:

$$\text{False Alarm(FA) Rate } = \frac{1}{m_-} \sum_{i \in \mathbb{D}_-} \mathbb{1}(\hat{y}^{(i)} \neq -1), \quad \text{Missed Detection (MD) Rate} = \frac{1}{m_+} \sum_{i \in \mathbb{D}_+} \mathbb{1}(\hat{y}^{(i)} \neq 1)$$

In this section, we will use test data set to evaluate the performance of the classifier. Before adjusting the weight (all weight=1) the classifier shown below(with the training dataset):

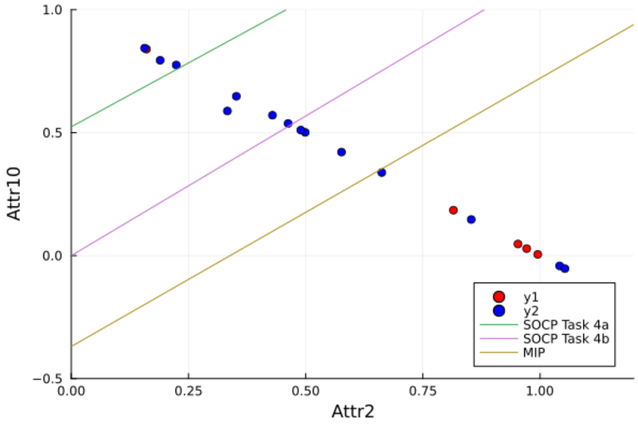

Figure 8: Classifier before adjusting the weight

$$\text{Training FA/MD rate of SOCP Task 4a} = (0.35714285714285715, 0.0)$$
$$\text{Test FA/MD rate of SOCP Task 4a} = (0.4897959183673469, 0.45454545454545453)$$
$$\text{Training FA/MD rate of SOCP Task 4b} = (0.21428571428571427, 0.16666666666666666)$$
$$\text{Test FA/MD rate of SOCP Task 4b} = (0.32653061224489793, 0.8181818181818182)$$
$$\text{Training FA/MD rate of MIP Task 4c} = (0.2857142857142857, 0.16666666666666666)$$
$$\text{Test FA/MD rate of MIP Task 4c} = (0.2857142857142857, 0.6363636363636364)$$

After adjusting the weights (weight1 = 1, weight2 = 1.332, weight3 = 0.8)(with the training dataset):

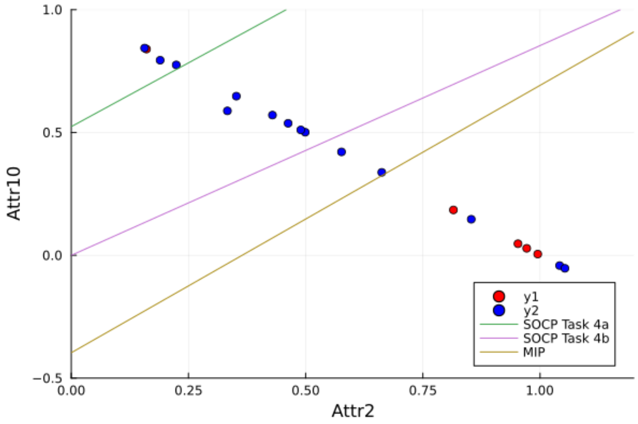

Figure 9: Classifier after adjusting the weight

$$\text{Training FA/MD rate of SOCP Task 4a} = (0.35714285714285715, 0.0)$$
$$\text{Test FA/MD rate of SOCP Task 4a} = (0.4897959183673469, 0.45454545454545453)$$
$$\text{Training FA/MD rate of SOCP Task 4b} = (0.35714285714285715, 0.16666666666666666)$$
$$\text{Test FA/MD rate of SOCP Task 4b} = (0.40816326530612246, 0.5454545454545454)$$
$$\text{Training FA/MD rate of MIP Task 4c} = (0.21428571428571427, 0.16666666666666666)$$
$$\text{Test FA/MD rate of MIP Task 4c} = (0.2653061224489796, 0.6363636363636364)$$

With adjusting the weighting, we successfully decrease the error rate of the classifier.

4 Competitive Task

In this section, we will consider the full dataset and utilize all the available attributes to detect bankruptcy. The objectives are to find a classifier with the best training / testing error and the sparsest feature selection. The requirements are as follow: (1) the classifier has to be found using a custom-made iterative algorithm such as projected gradient descent for solving an optimization problem of the form:

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \quad \hat{f}(w,b) \quad s.t \quad (w,b) \in X,$$

where $\hat{f}(w,b)$ is be built using the training dataset and $X \subseteq \mathbb{R}^d \times \mathbb{R}$ The objective function $\hat{f}(w,b)$:

$$\hat{f}(w,b) = \frac{1}{m} \sum_{i=1}^{m} \ell_i log(1 + \exp(-y^{(i)}((x^{(i)})^\top w + b)))$$

Constraint:

$$X = \{w \in \mathbb{R}^d, \ b \in \mathbb{R} : \ |b| + \sum_{i=1}^{d} |w_i| \leq R_1\}$$

We will implement an iterative algorithm: projected gradient decent(PGD) method, using a diminishing step size as $\gamma = 0.88$ to solve the optimization problem. PGD method is to iteratively refine a solution by taking steps in the direction of the gradient(i.e. the steepest ascent) of a function. And then projecting back onto the constraint set whenever the gradient step might lead the solution out of the set. This ensure the solution remain within the constraint.

Initially, we first set up a $weights = 1.88$, $R = 0.8$, $\gamma = 0.88$, and the max iterative number respectively. Then use PGD method, the pseudocode are as follow:

```
**Input:** $\boldsymbol{x}^{(0)} \in X$, constant step size $\gamma \leq 0$, max, iteration number $K_{max}$.
**For** $k = 0, ..., K_{max}$
    $\boldsymbol{x}^{k+1} = Proj_X\{\boldsymbol{x}^{(k)} - \gamma \nabla \hat{f}(\boldsymbol{x}^k)\}$
**End For**
```

For the projection operator, pseudocode are as follow

```
**Input**: $x \in \mathbb{R}^d, R > 0$
vector $u = abs.(x)$
vector $v = sort.(u)$ // sorted vector, i.e. $|v_1| \geq ... \geq |v_d|$
**For** $j = 1, ...d$
    **If** $v_j - \frac{1}{j}(\sum_{r=1}^{j} v_r - R) \leq 0$, **Then** set $j_{sv} = j - 1$ and break the for loop
Set $\theta = \frac{1}{j}(\sum_{r=1}^{j_{sv}} v_r - R)$
**Return** $\hat{x}$ such that $\hat{x_i} = sign \max 0, |x_i| - \theta$ for $i = 1, 2, ..d$
```

Using the program, we can get the graph of the objective value:

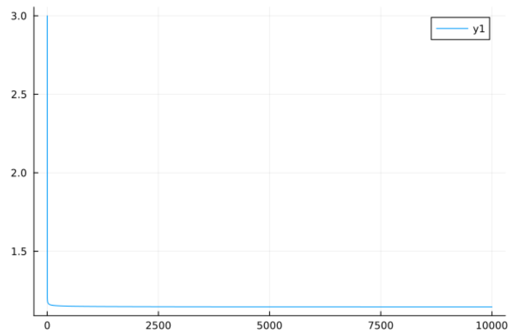

Figure 10: Objective Value

Then we will use the F1 score which is a common metric to evaluate the classifier performance:

$$F_1 = \frac{2(1 - P_{MD})}{2(1 - P_{MD}) + P_{FA} + P_{MD}}$$

Moreover, the number of non-zero elements in (w, b) will be calculated according to the normalized version of latter, which will implement in the program.

$$(\#\text{non-zero elements in } w, b) = \mathbb{1}(|\frac{|b|}{|b| + \sum_{j=1}^{d} |w_j|}| \geq 0.01) + \sum_{i=1}^{d} \mathbb{1}(|\frac{|w_i|}{|b| + \sum_{j=1}^{d} |w_j|}| \geq 0.01)$$

By calculating the F1 score for the training dataset and test dataset, and the non-zero element ,we can get the socre :

$$\text{train F1} = 0.6742966060369651$$

$$\text{test F1} = 0.6637718300383382$$

$$\text{no of non-zeros} = 3$$

After all, we have finish implementing the SVM and using PGD to optimize.

5 Conclusions

In this project, the SVM classifier is successfully implemented, optimized using the Projected Gradient Method. Which aim to predict to predict bankruptcy among Polish companies based on 64 performance indicators from 2000 to 2012. However, there are some uncovered challenges, such as tuning the SVM parameter. In the future, other model can be implemented to find the optimal parameter, such as Bayesian Optimization. Also, alternative machine learning algorithms can be used to compare the performance, in order to improve the performance.

Appendix A

A Dataset Description Here is the list of all the 64 features collected in the Bankruptcy dataset:
Attr1 net profit / total assets
Attr2 total liabilities / total assets
Attr3 working capital / total assets
Attr4 current assets / short-term liabilities
Attr5 [(cash + short-term securities + receivables - short-term- liabilities)/ (operating expenses - depreciation)] * 365
Attr6 retained earnings / total assets
Attr7 EBIT / total assets
Attr8 book value of equity / total liabilities
Attr9 sales / total assets
Attr10 equity / total assets
Attr11 (gross profit + extraordinary items + financial expenses) / total assets
Attr12 gross profit / short-term liabilities
Attr13 (gross profit + depreciation) / sales
Attr14 (gross profit + interest) / total assets
Attr15 (total liabilities * 365) / (gross profit + depreciation)
Attr16 (gross profit + depreciation) / total liabilities
Attr17 total assets / total liabilities
Attr18 gross profit / total assets
Attr19 gross profit / sales
Attr20 (inventory * 365) / sales
Attr21 sales (n) / sales (n-1)
Attr22 profit on operating activities / total assets
Attr23 net profit / sales
Attr24 gross profit (in 3 years) / total assets
Attr25 (equity - share capital) / total assets
Attr26 (net profit + depreciation) / total liabilities
Attr27 profit on operating activities / financial expenses
Attr28 working capital / fixed assets
Attr29 logarithm of total assets
Attr30 (total liabilities - cash) / sales
Attr31 (gross profit + interest) / sales
Attr32 (current liabilities * 365) / cost of products sold
Attr33 operating expenses / short-term liabilities
Attr34 operating expenses / total liabilities
Attr35 profit on sales / total assets
Attr36 total sales / total assets
Attr37 (current assets - inventories) / long-term liabilities
Attr38 constant capital / total assets
Attr39 profit on sales / sales
Attr40 (current assets - inventory - receivables) / short-term liabilities
Attr41 total liabilities / ((profit on operating activities + depreciation) * (12/365))
Attr42 profit on operating activities / sales
Attr43 rotation receivables + inventory turnover in days
Attr44 (receivables * 365) / sales
Attr45 net profit / inventory
Attr46 (current assets - inventory) / short-term liabilities
Attr47 (inventory * 365) / cost of products sold
Attr48 EBITDA (profit on operating activities - depreciation) / total assets
Attr49 EBITDA (profit on operating activities - depreciation) / sales
Attr50 current assets / total liabilities
Attr51 short-term liabilities / total assets
Attr52 (short-term liabilities * 365) / cost of products sold)
Attr53 equity / fixed assets
Attr54 constant capital / fixed assets
Attr55 working capital
Attr56 (sales - cost of products sold) / sales

Attr57 (current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
Attr58 total costs /total sales
Attr59 long-term liabilities / equity
Attr60 sales / inventory
Attr61 sales / receivables
Attr62 (short-term liabilities *365) / sales
Attr63 sales / short-term liabilities
Attr64 sales / fixed assets