CS 6350: Machine Learning – Homework 5

Clinton Fernandes, u1016390 November 15, 2016

1 Warm up: Margins

1.

x_1	x_1	XOR label	x_1x_2
-1	-1	–	1
-1	1	+	-1
1	-1	+	-1
1	1	–	1

- The boolean table on the left shows the values of x_1, x_2, the XOR labels and values of x_1x_2. The figure below(on the left) shows the transformed euclidean space in (x_1, x_1x_2) with the points plotted(with labels), while the figure on the right shows the original euclidean space in (x_1, x_1). The hyperplane is the thick line.

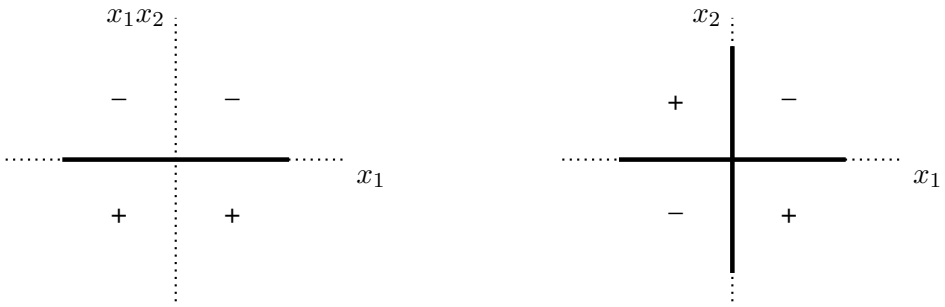

Figure 1: Transformed space (x_1, x_1x_2) on left and Original space (x_1, x_2) on the right.

The maximal margin is found to be 1 (distance from the hyperplane and the closest point). The hyperplane is represented by the thick line $x_1x_2 = 0$ in the transformed space (x_1, x_1x_2). Back in the original euclidean space (as shown by the figure on the right), the hyperplane $x_1x_2 = 0$ represents two lines along the axes x_1 and x_2 (hyperbola with two asymptotes).

2. (a) $D_1 = \{x_1, x_2, x_3, x_5, x_7\}$ $D_2 = \{x_1, x_5, x_6, x_8\}$ $D_3 = \{x_3, x_4, x_5, x_7\}$

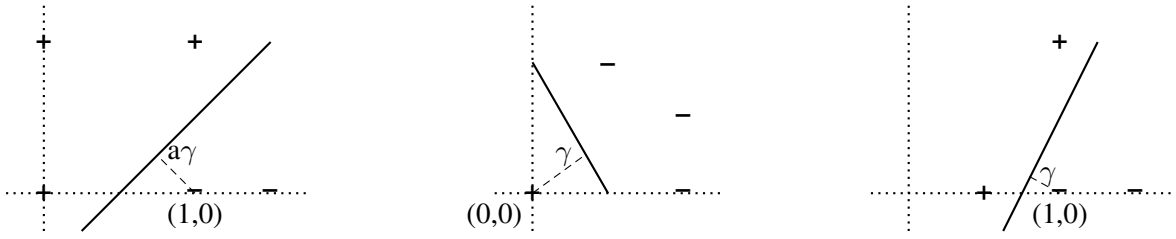

Figure 2: diagrams for dataset D_1 (left), dataset D_2 (center) and dataset D_3 (right)

For the dataset $D_1 = \{x_1, x_2, x_3, x_5, x_7\}$ (left figure):
The equation of the hyperplane that separates the positive and the negative labels of the data was

found to be $x + y - 0.5 = 0$. The closest points to this line were (1,0) or (1,1). The margin (distance between the point (1,1) and the hyperplane) was computed using the distance formula $\frac{ax+by+c}{\sqrt{a^2+b^2}} = \frac{1}{2\sqrt{2}} = 0.3535$.

For the dataset $D_2 = \{x_1, x_5, x_6, x_8\}$ (center figure):
The equation of the hyperplane that separates the positive and the negative labels was found to be: $\sqrt{3}x + y - \frac{\sqrt{3}}{2} = 0$. The closest points to the hyperplane were (0,0) or (1,0) or (0.5,$\sqrt{3}/2$). To get the margin, distance from (0,0) and the hyperplane was computed using the equation $\frac{ax+by+c}{\sqrt{a^2+b^2}}$. The margin was found to be $\sqrt{3}/4 = 0.4330$

For the dataset $D_3 = \{x_3, x_4, x_5, x_7\}$ (right figure):
The equation of the hyperplane was found to be $2x - y - 1.5 = 0$. The margin of the data using the distance between the closest point (1,0) and the hyperplane was found to be $1/\sqrt{5} = 0.2236$

(b) The perceptron mistake bound for any dataset is defined by $(R/\gamma)^2$.
Here, R is the farthest point from the origin and γ is the margin of the data.
For D_1, $R = 1.5$ and $\gamma = 0.3536$. So mistake bound $(R/\gamma)^2$ = 18
For D_2, $R = 1.118$ and $\gamma = 0.433$. So mistake bound $(R/\gamma)^2$ = 6.667
For D_3, $R = 1.5$ and $\gamma = 0.2236$. So mistake bound $(R/\gamma)^2$ = 45.0027
The dataset D_3 has the greatest mistake bound.

(c) In terms of ease of learning: $D_2 > D_1 > D_3$
The mistake bound signifies the finite number of mistakes that the algorithm will make on the training set until it finds the separating hyperplane. Lower is the mistake bound, the algorithm will require lesser number of examples in order to guarantee that the error will be less than required with high confidence and hence that particular dataset is easier to learn than a dataset on which the mistake bound is higher.

2 Kernels

1. (a) Proof: If $K_1(\mathbf{x}, \mathbf{z})$ and $K_2(\mathbf{x}, \mathbf{z})$ are kernels, then prove that $K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z})$ is a kernel.
let ϕ_1 be the feature map(transformation) for $K_1(\mathbf{x}, \mathbf{z})$ and let ϕ_2 be the feature map for $K_2(\mathbf{x}, \mathbf{z})$.
let $f_i(\mathbf{x})$ be the ith feature value of the feature map ϕ_1 and let $g_j(\mathbf{x})$ be the jth feature value of ϕ_2.

A valid kernel represents an inner product. Using this property for $K_1(\mathbf{x}, \mathbf{z})$ and $K_2(\mathbf{x}, \mathbf{z})$.

$$\begin{aligned} K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z}) &= (\phi_1(\mathbf{x})^T \cdot \phi_1(\mathbf{z}))(\phi_2(\mathbf{x})^T \cdot \phi_2(\mathbf{z})) \\ &= \left(\sum_{i=1}^{\infty} f_i(\mathbf{x}) f_i(\mathbf{z})\right)\left(\sum_{j=1}^{\infty} g_j(\mathbf{x}) g_j(\mathbf{z})\right) \\ &= \sum_{i,j} f_i(\mathbf{x}) f_i(\mathbf{z}) g_j(\mathbf{x}) g_j(\mathbf{z}) \\ &= \sum_{i,j} (f_i(\mathbf{x}) f_i(\mathbf{z}))\ (g_j(\mathbf{x}) g_j(\mathbf{z})) \end{aligned} \tag{1}$$

We can define a feature map ϕ_3 with its $\langle i, j \rangle$th feature as $h_{i,j}(\mathbf{x})$, represented by:

$$h_{i,j}(\mathbf{x}) = f_i(\mathbf{x}) g_j(\mathbf{x}) \tag{2}$$

Then we have

$$\begin{aligned} K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z}) &= \sum_{i,j} h_{i,j}(\mathbf{x}) h_{i,j}(\mathbf{z}) \\ &= \phi_3(\mathbf{x})^T \cdot \phi_3(\mathbf{z}) \end{aligned} \tag{3}$$

Here, the inner product sums over the pairs $\langle i, j \rangle$. Since the number of such pairs is countable, we can enumerate the pairs in a linear sequence to get $\phi_3(\mathbf{x})$.

As we have shown that $K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z})$ can be represented as an inner product, therefore it is a valid kernel.

(b) Proof: If P is any polynomial with positive coefficients, show that $K(\mathbf{x}, \mathbf{z}) = P(K_1(\mathbf{x}, \mathbf{z}))$ is a valid kernel

We can begin by showing $K(\mathbf{x}, \mathbf{z}) = \alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z})$ is a valid kernel and use the conclusion from the previous question.

First, consider the term $\alpha K_1(\mathbf{x}, \mathbf{z})$, and prove that it represents a valid kernel.

let ϕ_1 be a feature map for the kernel $K_1(\mathbf{x}, \mathbf{z})$. Then,

$$\begin{aligned} K_1(\mathbf{x}, \mathbf{z}) &= \phi_1(\mathbf{x})^T \cdot \phi_1(\mathbf{z}) \\ \alpha K_1(\mathbf{x}, \mathbf{z}) &= \alpha\ \phi_1(\mathbf{x})^T \cdot \phi_1(\mathbf{z}) \\ &= (\sqrt{\alpha}\phi_1(\mathbf{x}))^T \cdot (\sqrt{\alpha}\phi_1(\mathbf{z})) \end{aligned} \tag{4}$$

$(\sqrt{\alpha}\ \phi_1(\mathbf{x}))$ itself represents a feature map(transformation) and be represented as, $\phi_2(\mathbf{x}) = \sqrt{\alpha}\phi_1(\mathbf{x})$

$$\alpha K_1(\mathbf{x}, \mathbf{z}) = \phi_2(\mathbf{x})^T \cdot \phi_2(\mathbf{z}) \tag{5}$$

$\alpha K_1(\mathbf{x}, \mathbf{z})$ is represented as an inner product, hence it is a valid kernel

In a similar way, $\beta K_2(\mathbf{x}, \mathbf{z})$ can also be shown to be a valid kernel.

Now, let us represent the kernel $\alpha K_1(\mathbf{x}, \mathbf{z}) = K_1'(\mathbf{x}, \mathbf{z})$

and the kernel $\beta K_2(\mathbf{x}, \mathbf{z}) = K_2'(\mathbf{x}, \mathbf{z})$

$$\alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z}) = K_1'(\mathbf{x}, \mathbf{z}) + K_2'(\mathbf{x}, \mathbf{z}) \tag{6}$$

Now, we need to show that $K_1'(\mathbf{x}, \mathbf{z}) + K_2'(\mathbf{x}, \mathbf{z})$ represents a valid kernel.

From mercer's condition, for a valid kernel K, for every finite set $\{x_1, x_2, .. \}$, for any choice of real valued $c_1, c_2, ..$, we have:

$$\sum_i \sum_j c_i c_j K_1'(\mathbf{x}_i, \mathbf{x}_j) =\geq 0 \tag{7}$$

From equation 6 and 7

$$\begin{aligned} K_1'(\mathbf{x}, \mathbf{z}) + K_2'(\mathbf{x}, \mathbf{z}) &= \sum_i \sum_j c_i c_j K_1'(\mathbf{x}_i, \mathbf{x}_j) + \sum_i \sum_j c_i c_j K_2'(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \\ &= \sum_i \sum_j c_i c_j \left(K_1'(\mathbf{x}_i, \mathbf{x}_j) + K_2'(\mathbf{x}_i, \mathbf{x}_j) \right) \geq 0 \end{aligned} \tag{8}$$

As we know that $k_1'(\mathbf{x}, \mathbf{z})$ and $K_2'(\mathbf{x}, \mathbf{z})$ are valid kernels, and that a valid kernel represents a symmetric matrix, the addition of two symmetric matrices is a symmetric matrix.

And, $K_1'(\mathbf{x}, \mathbf{z}) + K_2'(\mathbf{x}, \mathbf{z})$ can be represented as $K_3(\mathbf{x}, \mathbf{z})$ which is a symmetric matrix.

From equation 8

$$K_1'(\mathbf{x}, \mathbf{z}) + K_2'(\mathbf{x}, \mathbf{z}) = \sum_i \sum_j c_i c_j \left(K_3(\mathbf{x}_i, \mathbf{x}_j) \right) \geq 0 \tag{9}$$

From mercers rule, $K_3(\mathbf{x}, \mathbf{z})$ is a valid kernel and hence $K_1'(\mathbf{x}, \mathbf{z}) + K_2'(\mathbf{x}, \mathbf{z})$ represents a valid kernel.

From the above statement and equation 6, $\alpha K_1(\mathbf{x}, \mathbf{z}) + \beta K_2(\mathbf{x}, \mathbf{z})$ represents a valid kernel.

This means that the sum of two kernels with positive coefficients is a valid kernel.

Now, $P(K_1(\mathbf{x}, \mathbf{z})) = a_n K_1(\mathbf{x}, \mathbf{z})^n + a_{n-1} K_1(\mathbf{x}, \mathbf{z})^{n-1} + \ldots + a_1 K_1(\mathbf{x}, \mathbf{z}) + a_0$

$P(K_1(\mathbf{x}, \mathbf{z}))$ represents a sum of products of kernels with positive coefficients.

We have proved in previous question (Q.2.1.a) that the product of kernels is a kernel and (in the previous statements) that the sum of kernels with with positive coefficients is a kernel. Hence $P(K_1(\mathbf{x}, \mathbf{z}))$ is a valid kernel.

2. Prove that $K(\mathbf{x}, \mathbf{z}) = 15 \left(\mathbf{x}^T \mathbf{z} \right)^2 \exp\left(-||\mathbf{x} - \mathbf{z}||^2 \right)$ is a valid kernel,

   Rewrite $K(\mathbf{x}, \mathbf{z}) = 15 \left(\mathbf{x}^T \mathbf{z} \right) \left(\mathbf{x}^T \mathbf{z} \right) \exp\left(-||\mathbf{x} - \mathbf{z}||^2 \right)$

   Consider the term $\left(\mathbf{x}^T \mathbf{z} \right)$

   This is a product of two features $\mathbf{x}$ and $\mathbf{z}$. If there was a feature map of these features then it would be $\phi(\mathbf{x}) = \mathbf{x}$. Then, we can write $\left(\mathbf{x}^T \mathbf{z} \right) = \phi(\mathbf{x})^T \cdot \phi(\mathbf{z})$. This is an inner product of two feature maps, and is the simple linear kernel. Name this as $K_1(\mathbf{x}, \mathbf{z}) = \left(\mathbf{x}^T \mathbf{z} \right) = \phi(\mathbf{x})^T \cdot \phi(\mathbf{z})$.

Now, consider:

$$\begin{aligned}\exp\left(-||\mathbf{x}-\mathbf{z}||^2\right) &= \exp\left(-||\mathbf{x}||^2 - ||\mathbf{z}||^2 + 2\mathbf{x}^T\mathbf{z}\right) \\ &= \exp\left(-||\mathbf{x}||^2\right)\ \exp\left(-||\mathbf{z}||^2\right)\ \exp\left(2\mathbf{x}^T\mathbf{z}\right) \\ &= \exp\left(-\mathbf{x}^T\mathbf{x}\right)\ \exp\left(-\mathbf{z}^T\mathbf{z}\right)\ \exp\left(2\mathbf{x}^T\mathbf{z}\right)\end{aligned} \tag{10}$$

Consider the term: $\exp\left(2\mathbf{x}^T\mathbf{z}\right)$

We have the taylor series expansion: $exp(x) = lim_{n\to\infty}(1 + x + \frac{x^2}{2!}... + \frac{x^n}{n!}) = \sum_{n=0}^{\infty}\frac{x^n}{n!}$

$$\begin{aligned}\exp\left(2\mathbf{x}^T\mathbf{z}\right) &= 1 + 2\mathbf{x}^T\mathbf{z} + \frac{(2\mathbf{x}^T\mathbf{z})^2}{2!}... + \frac{(2\mathbf{x}^T\mathbf{z})^n}{n!} \\ &= a_0 + a_1\mathbf{x}^T\mathbf{z} + a_2(\mathbf{x}^T\mathbf{z})^2... + a_n(\mathbf{x}^T\mathbf{z})^n\end{aligned} \tag{11}$$

$a_0, a_1, ...$ represent the coefficients.

The taylor series expansion of $exp(2\mathbf{x}^T\mathbf{z})$ represents a kernel with (an infinite set of) features corresponding to polynomial terms, because it is already proven in question 2.1.b that polynomial over a kernel that is constructed using positive coefficients is a kernel.

Now consider the terms $\exp\left(-\mathbf{x}^T\mathbf{x}\right)\ \exp\left(-\mathbf{z}^T\mathbf{z}\right)$

We can represent the feature transformation $\phi(\mathbf{x}) = \exp\left(-\mathbf{x}^T\mathbf{x}\right)$

where, $\phi(\mathbf{x})$ contains only one single feature $\mathbf{x}$.

Hence $\exp\left(-\mathbf{x}^T\mathbf{x}\right)\ \exp\left(-\mathbf{z}^T\mathbf{z}\right) = \phi(\mathbf{x})^T \cdot \phi(\mathbf{x})$

$\exp\left(-\mathbf{x}^T\mathbf{x}\right)\ \exp\left(-\mathbf{z}^T\mathbf{z}\right)$ is represented as an inner product, hence it is a valid kernel.

We have proven independently that $\exp\left(2\mathbf{x}^T\mathbf{z}\right)$ and $\exp\left(-\mathbf{x}^T\mathbf{x}\right)\ \exp\left(-\mathbf{z}^T\mathbf{z}\right)$ are kernels. And the product of these two should also be a kernel (Q.2.1.a).

Then represent the kernel, $\exp\left(2\mathbf{x}^T\mathbf{z}\right)\exp\left(-\mathbf{x}^T\mathbf{x}\right)\ \exp\left(-\mathbf{z}^T\mathbf{z}\right) = K_2(\mathbf{x}, \mathbf{z})$

Now, we have $K(\mathbf{x}, \mathbf{z}) = 15\left(\mathbf{x}^T\mathbf{z}\right)^2\exp\left(-||\mathbf{x}-\mathbf{z}||^2\right) = 15\left(K_1(\mathbf{x}, \mathbf{z})\right)^2 K_2(\mathbf{x}, \mathbf{z})$

$\left(K_1(\mathbf{x}, \mathbf{z})\right)^2 K_2(\mathbf{x}, \mathbf{z})$ represents a product of kernels and so it is a kernel (Q.2.1.a), let this be $K_3(\mathbf{x}, \mathbf{z})$

$K(\mathbf{x}, \mathbf{z}) = 15\left(\mathbf{x}^T\mathbf{z}\right)^2\exp\left(-||\mathbf{x}-\mathbf{z}||^2\right) = 15K_3(\mathbf{x}, \mathbf{z})$

The right hand side is in the form $\alpha K(\mathbf{x}, \mathbf{z})$ which was proven to be a kernel in (Q.2.1.b equations 5-6)

Hence $K(\mathbf{x}, \mathbf{z}) = 15\left(\mathbf{x}^T\mathbf{z}\right)^2\exp\left(-||\mathbf{x}-\mathbf{z}||^2\right)$ is a valid kernel.

3. Prove that the Gaussian kernel, $K(\mathbf{x}, \mathbf{z}) = \exp(\frac{-\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2})$ can be written down as the inner product of an feature space with infinite dimension.

$$\begin{aligned}\exp(\frac{-\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}) &= \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2} + \frac{(\mathbf{x}^T\mathbf{z})^2}{\sigma^2} - \frac{\|\mathbf{z}\|^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{\|\mathbf{x}\|^2 + \|\mathbf{z}\|^2}{2\sigma^2}\right)\exp\left(\frac{\mathbf{x}^T\mathbf{z}}{\sigma^2}\right)\end{aligned} \tag{12}$$

Consider the term: $\exp\left(\frac{\mathbf{x}^T\mathbf{z}}{\sigma^2}\right)$

We have the taylor series expansion as: $exp(x) = lim_{n\to\infty}(1 + x + \frac{x^2}{2!}... + \frac{x^n}{n!}) = \sum_{n=0}^{\infty}\left(\frac{x^n}{n!}\right)$

$$\begin{aligned}\exp\left(\frac{\mathbf{x}^T\mathbf{z}}{\sigma^2}\right) &= \sum_{n=0}^{\infty}\left(\frac{\left(\frac{\mathbf{x}^T\mathbf{z}}{\sigma^2}\right)^n}{n!}\right) \\ &= \sum_{n=0}^{\infty}\left(\frac{(\mathbf{x}^T\mathbf{z})^n}{\sigma^{2n}n!}\right)\end{aligned} \tag{13}$$

Back in the previous equation 12, we have:

$$\begin{aligned}\exp(\frac{-\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}) &= \exp\left(-\frac{\|\mathbf{x}\|^2+\|\mathbf{z}\|^2}{2\sigma^2}\right)\sum_{n=0}^{\infty}\left(\frac{(\mathbf{x}^T\mathbf{z})^n}{\sigma^{2n}n!}\right) \\ &= \sum_{n=0}^{\infty}\left(\exp\left(-\frac{\|\mathbf{x}\|^2}{\sqrt{2}\sigma}\right)\frac{(\mathbf{x}^T)^n}{\sigma^{2n}\sqrt{n!}}\right)\left(\exp\left(-\frac{\|\mathbf{z}\|^2}{\sqrt{2}\sigma}\right)\frac{(\mathbf{z})^n}{\sigma^{2n}\sqrt{n!}}\right) \\ &= \sum_{n=0}^{\infty}\left(\frac{(\mathbf{x}^T)^n}{\sigma^{2n}\sqrt{n!}}\exp\left(-\frac{\|\mathbf{x}\|^2}{\sqrt{2}\sigma}\right)\right)\left(\frac{(\mathbf{z})^n}{\sigma^{2n}\sqrt{n!}}\exp\left(-\frac{\|\mathbf{z}\|^2}{\sqrt{2}\sigma}\right)\right)\end{aligned} \tag{14}$$

In the above expression, the Gaussian kernel is expressed as an inner product of feature space with infinite dimensions.

3 Experiments

3.1 Support Vector Machines

1. After Implementing SVM in handwriting dataset with C = 1 and γ_0 = 0.01
   Training Accuracy = 93.2%
   Training Accuracy = 91.4%

2. After running SVM madelon dataset and using 5-fold cross-validation to choose suitable parameters, the results obtained are shown in the table below.

Table 1: Results of cross validation with hyper-parameters C and γ_0

	C						
	-	2	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}
γ_0	0.1	50.00	56.45	55.65	54.8	53.85	54.5
	0.01	50.00	53.9	56.45	56.9	55.3	54.55
	0.001	52.55	52.2	52.7	55.8	54.55	51.85

The best accuracy was found to be 56.9% with C = $2^{-5} = 0.03125$ and $\gamma_0 = 0.01$

With these parameters, the Training accuracy = 61.1% and Test accuracy = 57.67%

Table 2: F1 score, Precision and Recall on classifiers

	Handwritten data (Q-3.1.1)		Madelon data (Q-3.1.2)	
	Training	Test	Training	Test
Precision	0.935	0.918	0.603	0.573
Recall	0.942	0.915	0.649	0.597
F1 score	0.938	0.916	0.625	0.584

3. It can be seen form table 2 that, for the the classifier from question 3.1.1, the F1 score, along with the precision and recall were found to be fairly good. However, on the classifier obtained from the question 3.1.2, the performance parameters, F1 score, precision and also recall were poor.

3.2 Ensemble of decision trees

1. After growing the N = 5 decision trees, a new dataset D consisting of transformed features was used to train the SVM meta-classifier.
   C = 1 $\gamma_0 = 0.01$ Number of epochs for SVM = 2
   Training Accuracy = 100.0%
   Test Accuracy = 96.3%

2. It was seen that the feature values of the madelon dataset were real valued and continuous. Hence, in order to use this dataset in the decision tree ensembles technique, the feature values had to be descritized using a feature transformation. This was done before giving the data to the ID3 algorithm.

   In the feature transformation, for each of the features:
   The mean(μ) and the standard deviation(σ) was calculated.
   Feature values greater than $(\mu + \sigma)$ were set to $(\mu + \sigma)$.
   Then, the feature values were replaced as: feature = feature/$((\mu + \sigma)/\sqrt{\sigma})$
   NOTE: due to feature transformation, the code takes some time to read and transform the data.

   (a) The method of ensembles of decision trees followed by SVM was implemented on the Madelon data, with C = 1 $\gamma_0 = 0.01$ number of samples (m) = 2000 # SVM epochs = 1
   For N = 10:
   Training Accuracy = 99.0%
   Test Accuracy = 62.67%
   For N = 30:
   Training Accuracy = 100.0%
   Test Accuracy = 65.33%
   For N = 100:
   Training Accuracy = 100.0%
   Test Accuracy = 69.33%

(b) Best N = 100

For Training:

Accuracy: 100.0%

Precision: 1.0

Recall: 1.0

F1 score: 1.0

For Test:

Accuracy: 69.33%

Precision: 0.653

Recall: 0.827

F1 score: 0.729