1 Decision Trees for Classification

$$Y = \text{sign}\left(\sum_{i=1}^{5} 0.9^i X_i + N(0,\sigma^2)\right), \qquad X_i = \begin{cases} -1, & \text{Prob} = 1/2 \\ 1, & \text{Prob} = 1/2 \end{cases} \text{ for } i = 1,2,\dots,15$$

1) $\sigma = 0.05$, training dataset size $= 5000$, test dataset size $= 500$. The following is the number of misclassifications on the training data and testing data as a function of minimum sample size to grow the tree. The $x-axis$ is minimum sample size the tree use and $y-axis$ is the number of misclassifications.

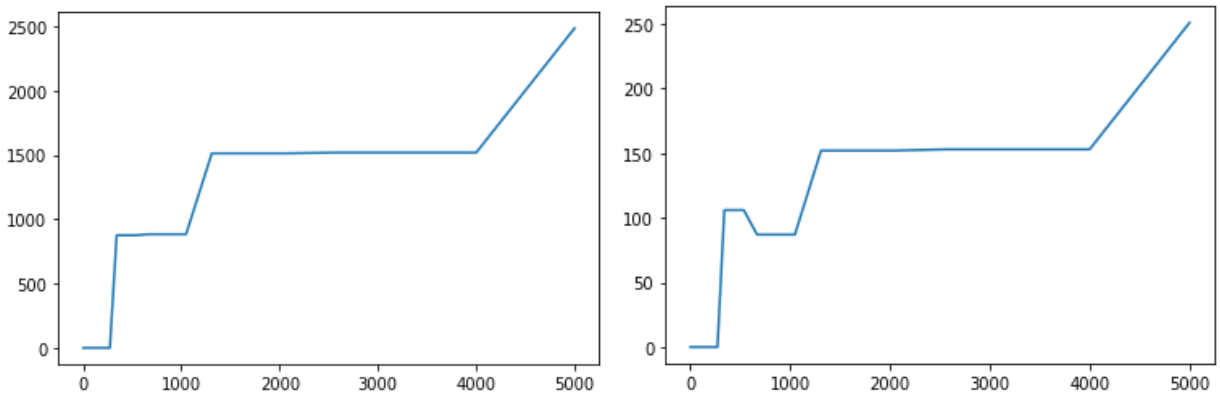

Figure 1: Training data (Left) and Testing data (Right)

The optimal size is 273, which is the maximum sample size that make the number of misclassifications lower to 0.

2) Repeat for a few times, the optimal sample size is varying between 218 and 273 when $\sigma = 0.05$. Thus, $s = 262$ by the 10 times trying.

```
finish trying  0 : the optimal size is  218
finish trying  1 : the optimal size is  273
finish trying  2 : the optimal size is  273
finish trying  3 : the optimal size is  273
finish trying  4 : the optimal size is  218
finish trying  5 : the optimal size is  273
finish trying  6 : the optimal size is  273
finish trying  7 : the optimal size is  273
finish trying  8 : the optimal size is  273
finish trying  9 : the optimal size is  273
the average optimal sample size is 262.0
```

Figure 2: running for acerage optimal sample size

3) Plot the misclassifications in function of σ. The following is the training and testing error on a decision tree grown to sample size $s = 262$ when σ changes, where $x - axis$ is standard deviation σ, and $y - axis$ is the number of misclassifications.

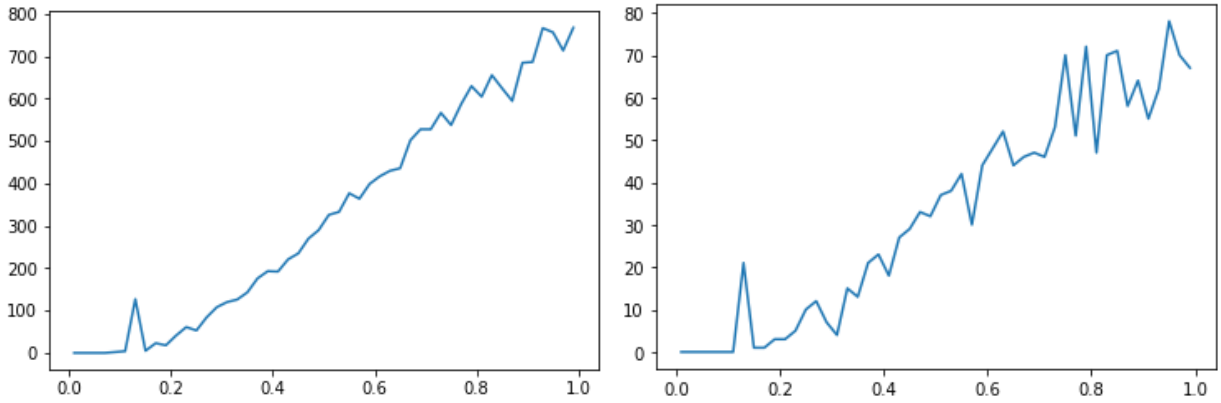

Figure 3: Training data (Left) and Testing data (Right)

When the variance of the random noise terms gets higher, the tree become less accurate, that is, lower the effectiveness of the tree.

4) Plot the number of times irrelevant features show up in the tree.

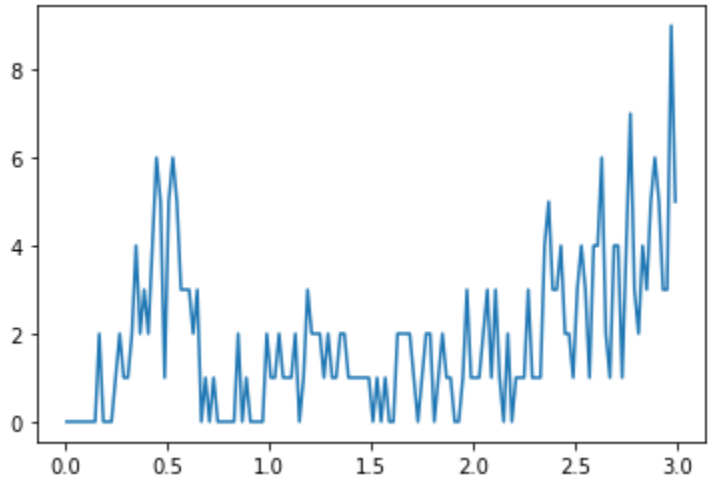

Figure 4: number of irrelevant features in decision tree

For $\sigma \in [0,1]$, the influence pattern is that when σ is increasing, the irrelevant features appear more frequently and reach the max at around $\sigma = 0.5$, then start decreasing until $\sigma = 1$. For $\sigma \in (1,3)$, the influence is getting higher when σ increase.

2 Logistic Regression

1) $\sigma = 0.05$, training dataset size $= 5000$, test dataset size $= 500$. Using fraction of $Y = +1$ in the terminal node. The following is the logistic error of the decision tree on training and testing data as a function of the minimum sample size that used to grow the tree.

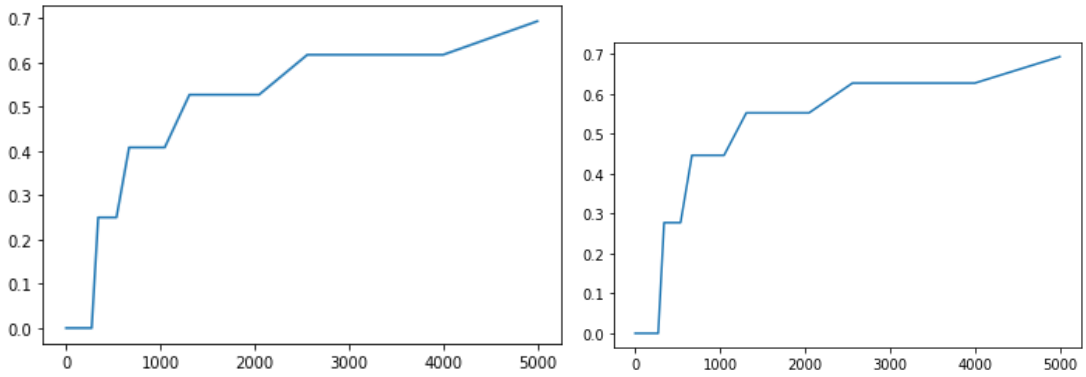

Figure 5: Training data (Left) and Testing data (Right)

The optimal sample size to grow the tree is 273, which is consist with the previous section.

2) $F(\underline{x}) = \sigma(w_0 + w_1x_1 + w_2x_2 + \cdots + w_{15}x_{15}) = 1/1 + \exp(-\underline{w} \cdot \underline{x})$. The following is the plot of the logistic error as a function of time as fit the model. $x - axis$ is time, $y - axis$ is error.

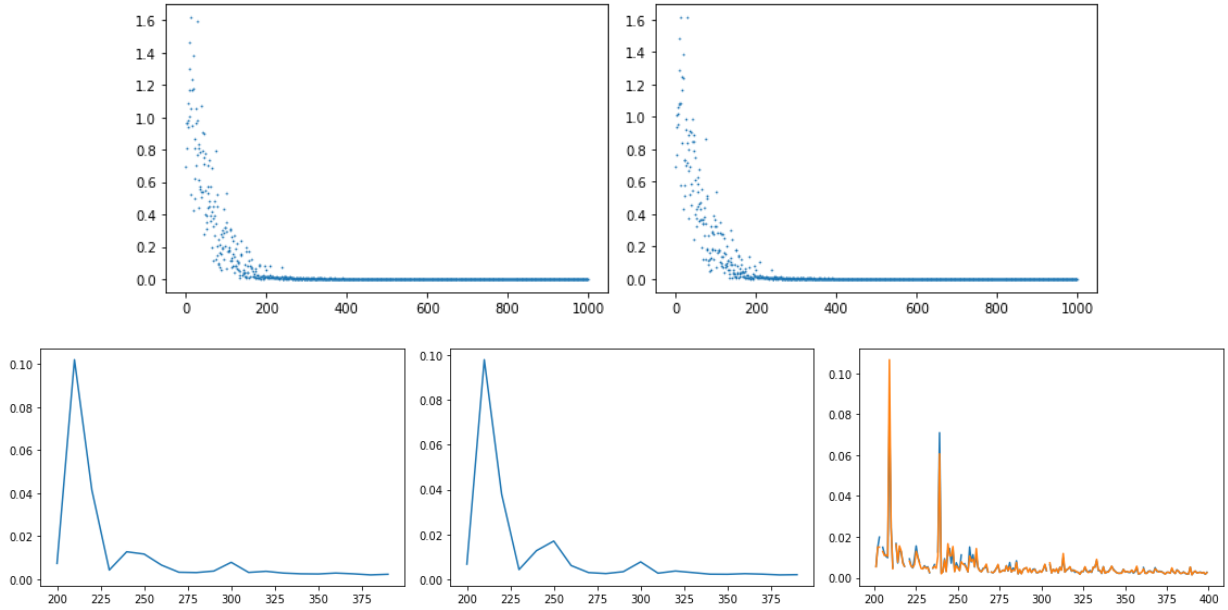

Figure 6: Training data (Left) and Testing data (Right)

Results for 1000 runs:

$\underline{w} = [w_0\ w_1\ w_2 \dots w_{15}] =$

```
 0.2168268   6.93567248   7.06529778   6.96890307  6.76582385   6.79271896
 0.06339229  0.05856282  -0.09993894   0.51284064  0.15745878  -0.02062972
-0.10500103 -0.3261003    0.23166924  -0.07288885
```

And the corresponding error is $\text{Error} = 0.0008556375337566522$

The fluctuation of the logistic error is large, but the overall trend is downward. Thus, the overfitting is not an issue here.

3) The decision tree model is better for modeling the data. Since the parameter $w_6, \dots, w_{15}$ is not so as close to 0, the decision tree is also better at minimizing the influence of irrelevant features.

3 Support Vector Machines

$(X_1, X_2), Y$ is

$$X = \begin{matrix}[-1,-1]\\[+1,-1]\\[-1,+1]\\[+1,+1]\end{matrix}, \qquad Y = \begin{matrix}-1\\+1\\+1\\-1\end{matrix}$$

Choose the Quadratic Kernel, which is the same as Polynomial Kernel with $d = 2, c = 1$. That is,

$$K(\underline{x}^i, \underline{x}^j) = (1 + \underline{x}^i \cdot \underline{x}^j)^2$$

To maximize the Dual SVM, we get that $\forall \alpha_i, \alpha_i = \frac{1}{8} = 0.125$, solve and get

$$\underline{w}^* = 0, \qquad \underline{b}^* = Y - 0 = [-1, +1, +1, -1]$$

$$Classifier(\underline{x}) = sign\left(\sum_i \alpha_i y_i K(\underline{x}^i, \underline{x}) + b\right)$$

The following is the plot of the classifier on the (X_1, X_2) plane, where

- $x - axis$ is X_1 and $y - axis$ is X_2
- Blue is the area for positive results, purple is the area for negative results, white is the line where the classifier is and has no classifications, which is consist by two lines $X_1 = 0$ and $X_2 = 0$.