Jessica Warren
STA 4634 Decision Tree Report
Arcene, Madelon and Gisette Datasets

Arcene Decision Tree Report

Decision Trees are a classification method commonly used in Machine Learning. The purpose of this assignment was to create a decision tree that would predict the labels for valid data. In order to predict the labels from the data set to know how accurately they match the given labels, I formulated a program in Python3. I used the package sklearn,to utilize the function score and the decision tree classifier.

Depth	1	2	3	4	5	6	7	8	9	10	11	12
Training error	.27	.14	.07	.03	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Valid Error	.34	.40	.38	.41	.38	.39	.42	.39	.36	.36	.39	.37

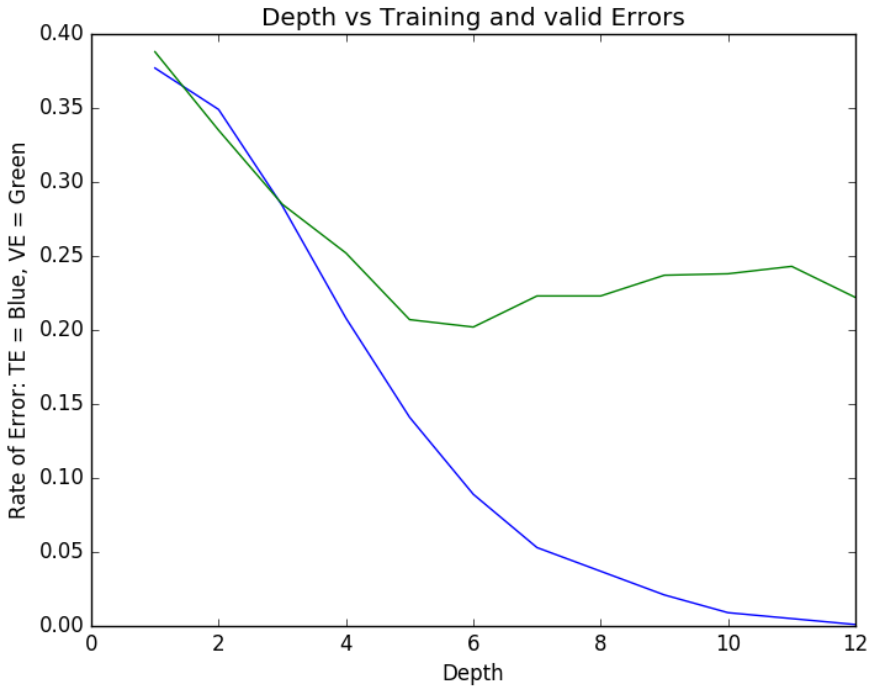
Jessica Warren
STA 4634 Decision Tree Report
Arcene, Madelon and Gisette Datasets

Analysis

The graph shows error, error shows the probability of making a mistake. The training curve in blue shows that as the depth of the tree increases, the misclassification error decreases. The Valid data error shown in green shows the concept of over fitting the data. The data at a depth of one predicts the labels with the least error with a score of .34. From 2 to 11 the error rate continues to fluctuate. At 12 the error rate again goes back to .34. This example of the data being simpler at 1 is an application of Occam's Razor. It is possible that because the design experiment in training had multiple different sources is why the data is very spurious.

Madelon Dataset Decision Tree Report

Depth	1	2	3	4	5	6	7	8	9	10	11	12
Training Errors	.38	.35	.29	.21	.001	.005	.009	.021	.037	.053	.089	.14
Valid Errors	.39	.35	.29	.25	.20	.22	.22	.24	.24	.24	.24	.22



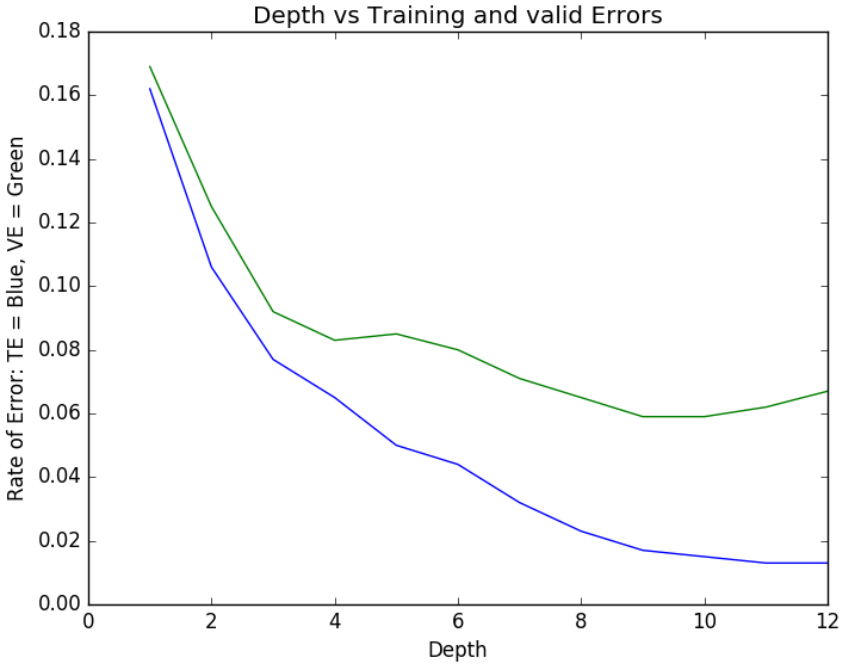
Jessica Warren
STA 4634 Decision Tree Report
Arcene, Madelon and Gisette Datasets

Analysis

The training curve in blue shows that as the depth of the tree increases, the misclassification error decreases. The Valid data error shown in green shows the concept of over fitting the data. The data is not as spurious as the Arcene data set but it does get over fit when the depth increases past 6.

Gisette

Depth	1	2	3	4	5	6	7	8	9	10	11	12
Training Errors	.162	.106	.077	.065	.05	.044	.032	.023	.017	.015	.013	.013
Valid Errors	.17	.13	.092	.083	.085	.08	.071	.065	.059	.059	.062	.067



Jessica Warren
STA 4634 Decision Tree Report
Arcene, Madelon and Gisette Datasets

Analysis

This data set has more data then Arcene and Madelon. It also shows the least error it its predictions. Over-fitting is found after the depth of 9. With out more knowledge concerning what the data values mean it is impossible to know why the valid data here is better predicted here rather then in Acene and Gisette. IThe gissete dataset is asking a question that has one of two answers is the pixels representing a four or a nine? t is possibble that because the dataset from Gisette was just meant to determine one out of two choices, that the data was predicted at a higher accuracy. Rather the other datasets although they have binary questions they are harder questions to answer and there is more randomization of data in the sets.