HW 2

YI NIAN
MPCS Machine learning, University of Chicago

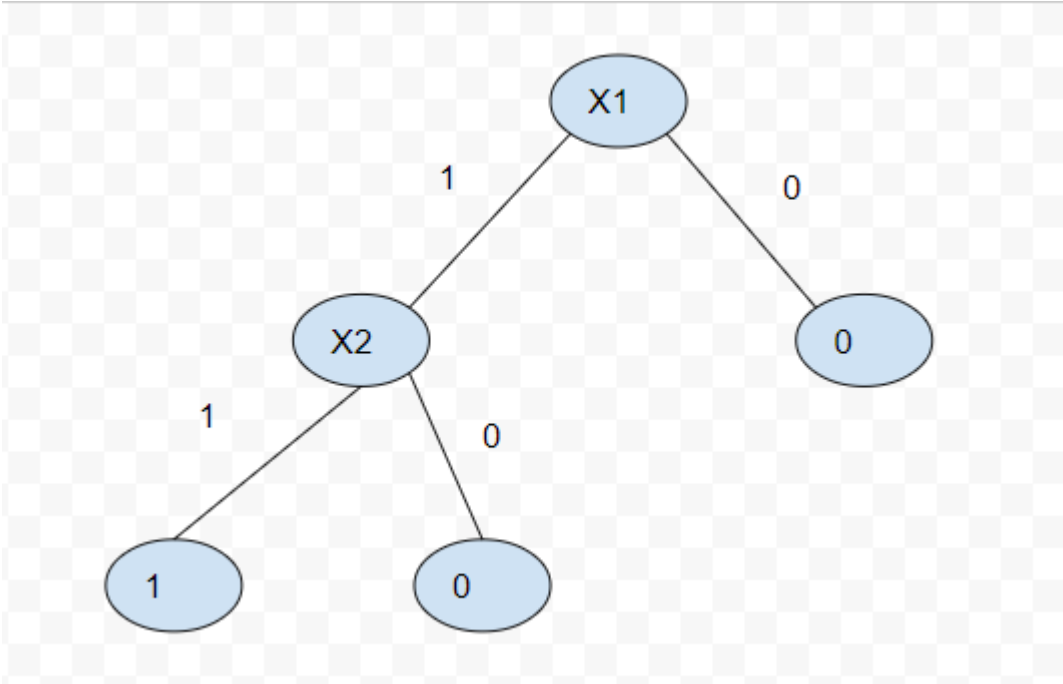
4-21

Collaborate with Hanqi Zhang on problem 1 and 3.

Problem 1:

1 18.13

One can transform every decision tree to decision list by pick positive outcomes' path as (Path1,1)–(Path2,1)...(Pathn,1)–(True,0). By doing this, we will get a decision list than has rules less than the number of leaves. In the worst case, we can write every path as conjunction of split nodes(conditions). In this way, we will have (leaf1,label1)–(leaf2,label2)—(leafn,labeln) at most number of leaves rules.



The above tree has 3 leaves. To show this in decision list, this is just (X1X2,1)–(True,0),which is less than the number of leaves.

2 18.14

3 a

In Boolean functions, for each possible input, we can think of all binary input as a conjunction(combination) p_i = $input_1, input_2.....input_n$. If the size of the test is unlimited, then we can write all possibility of inputs(p_1,p_2....p_n) with output label 1 and other inputs with label 0, as a decision list in this form:(p_1,1) – (p_2,1) – .....(p_n,1)–(True,0), which represent all the possibility of a Boolean function.

4 b

If we assign each leaf with label 1 when write it in DNF(disjunctive normal form), then we can use k literals to represent each path with depth k, it will be:

(path1)∪ (path2) ∪ .....(pathn), each path with length at most k.

Then the decision list will be:

(path1,1) –(path2,1)–....(pathn,1)–(True,0)

Problem 2:

5 a

True

6 b

False. The choice we make based on the information that 5 is the best among the training set. We cannot use the error of our best model to estimate the whole generalized error rate. There might be a model better or worse than max depth 5 and we cannot estimate this.

7 c

True.

This is the same as the statement 1 since we transform the randomness from "taking average of the sample" to random sampling.

Problem 3:

8 a

E[C_i| majority is 1] = number of 1 in S_i/number of examples in S_i = p

E[C_i| majority is 0] = 1-p

E[C_i] = E[C_i| majority is 1]*P(majority is 1) + E[C_i| majority is 0]*P(majority is 0)

E[C] = E[C_i] = p*P(majority is 1) + (1-p)*P(majority is 0)

Assume we have n = ((k-1)/k)*m samples, then

P(majority is 1 in T_i) = $\binom{n}{n}$ ·p^n+ $\binom{n}{n-1}$ ·p$^n(1-p)$+.... $\binom{n}{\lceil n/2 \rceil}$ ·p$^{\lceil n/2 \rceil}(1-p)^{\lceil n/2 \rceil}$

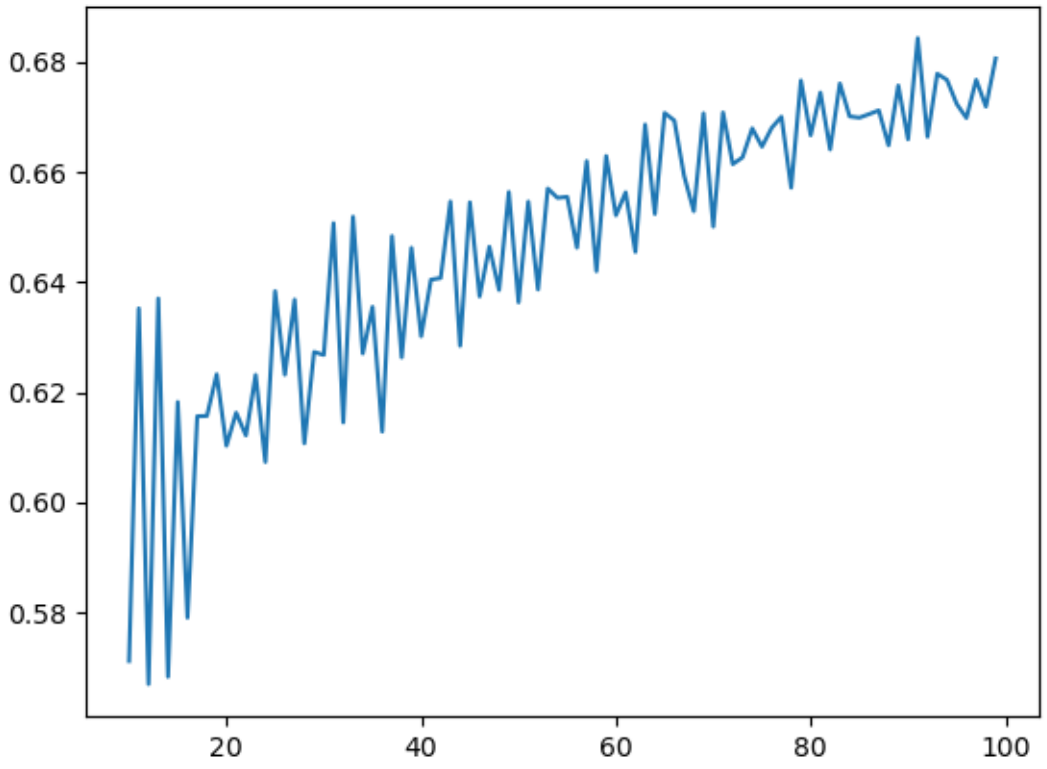
P(majority is 0 in T_i) = 1 - P(majority is 1 in T_i)

9 b

See python file

accuracy is around 0.65 for 1000 choices of D for example.

10 c



overall trending is increasing as the sample size increase.

11 d

As m goes to infinity, the proportion of 1 in the training set T will converge to p. In other words, the probability that majority of the training set is one, is actually converge to one. This means we will almost always predict 1 for the training set. Assume the indicator variable I_i denote whether the examples is in class 1 or not.

Then C = $\bar{I}_i$, the population average of the indicator.

And also, we have E($\bar{I}_i$) = p

according to LLN, P($|\bar{I}_i - E(I_i)| > \epsilon$) $\to$ 0 as m $\to \infty$ which implies to: P($|C - p| > \epsilon$) $\to$ 0 as m $\to \infty$

12 e

Assume we have m examples where half of them are 1 and half of them are 0 and we have k = 2 folds(Worst case).

During training, assume we pick all 1s as training set, then we will predict 1 during validate on all 0s, this leads to accuracy of 0.

On the other hand, we will predict 0 during validate on all 1s. this again leads to accuracy of 0.

Problem 4:

13 a

If you choose a to be 1, then there are n selections to choose b from 1 to n. Next, if we choose a to be 2, then there are (n-1) selections to choose b, which is from 2 to n. In this way, a can be chosen to be 1,2....n. And the total possibilities of (a,b) pairs are: (n+n-1+n-2+....1) = n(n+1)/2

In the same way, (c,d) pairs can also has n(n+1)/2 ways.

Therefore, (a,b),(c,d) can have $(n(n+1)/2)^2$ rectangles.

14 b

$n \geq 1/\epsilon(ln|H| + ln(1/\delta))$

$n \geq 1/0.1(ln(n(n+1)/2)^2 + ln(1/0.05))$

=10(2(ln(n(n+1))+ ln(20)))

=20ln($n^2 + n$)+10ln(20)

Problem 5:

Please see the code

Problem 6:

Some of the observations from question 6:

Part b

I use the 'most frequency' approach to impute the data since it does not make sense to impute some of the categorical variable use mean.

Adding additional boolean feature to indicate which of the rows had missing values

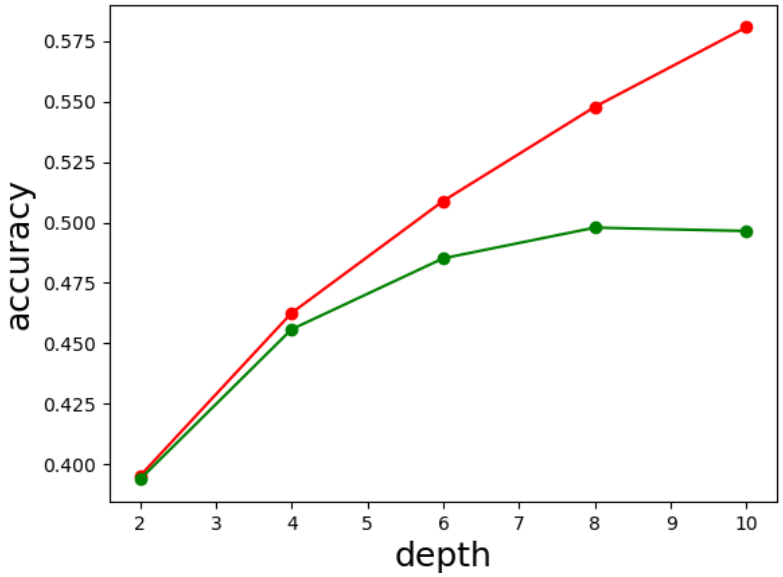
This approach might be helpful depends on the property on this data and also the model we use.

First, adding the boolean column will keep the information that which row is missing information. While this information might be correlated with the predicted data. So it might be helpful if we includes this information in a separated row. This new column might have small feature importance depends on the relation with predicted variable and the model.

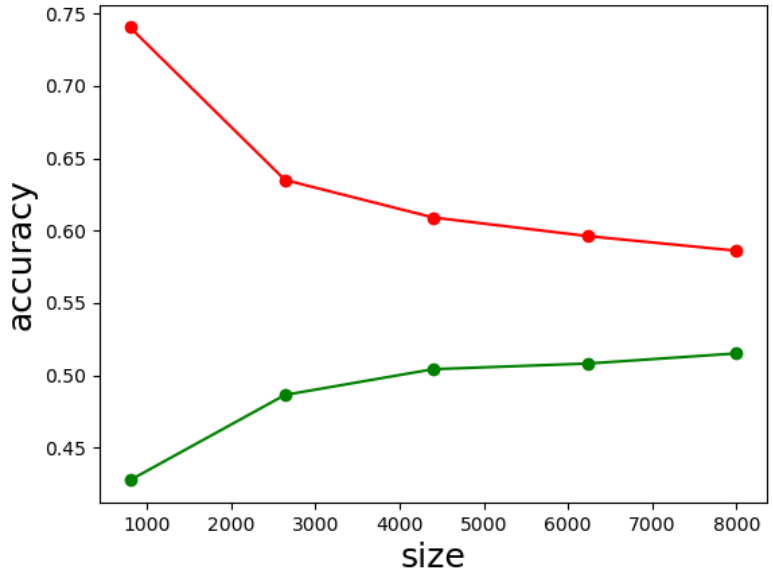
Sometimes, this information might be directed related to the variable that we need to predict. Then we will have data leakage problem in practice and we want to avoid this situation. But in the Kaggle competition, we actually want to capture such feature and take advantage of it.

Part d

Validation curve

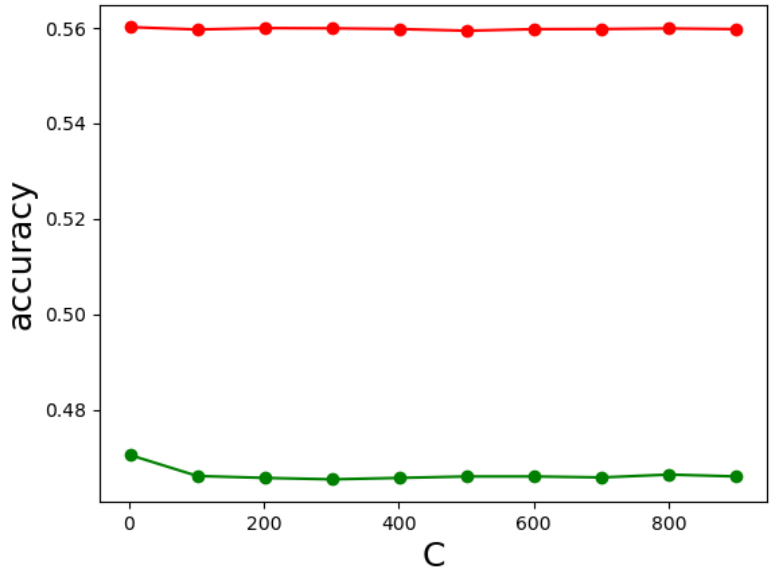


Learning Curve

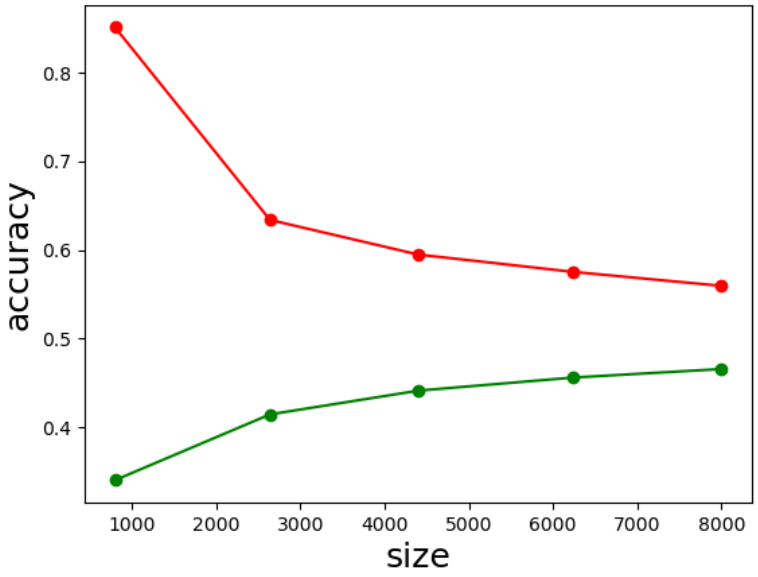


Part e

Validation curve



Learning Curve



Compare the learning curves for DecisionTreeClassifier and LogisticRegression and describe what you can learn from them.

Observation 1:

Both of the learning curves shows that it will be helpful if we adding more training data. Because the training curve and validation curve does not converge enough.As we get more data, the complexity of our model and the 'truth' model will become closer.

Observation 2:

The gap between the training curve and validation shows that we have a high variance problem. This might because of we have too many features compared to the training examples and this will cause overfitting. We can solve this by increase the data or reduce the features.

Observation 3:

Another approach to avoid this is to increase the regularization term for the logistic regression. But since we are not using the whole dataset because of the time issue, this might not be needed.