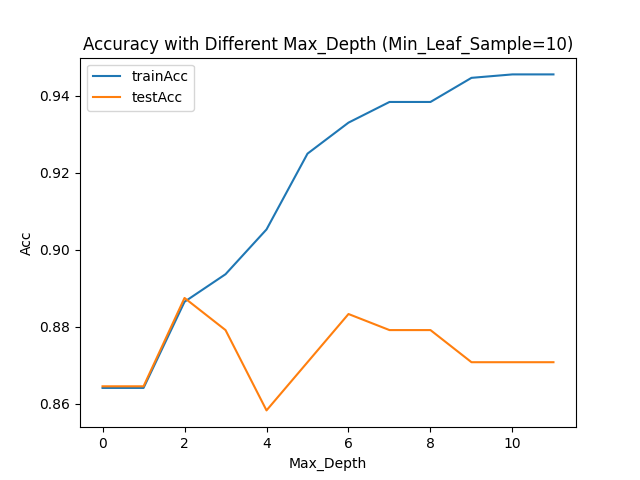
Q1

图表, 折线图

描述已自动生成c)

d) Time complexity of the train function: *O( d \* n \* 2^p)*

Because the maximum depth is *p* and decision tree is a binary tree, the train function needs to construct *2^p* nodes maximum. In each node construction, there is a double loop that traverse all possible splits in all attributes to calculate the information gain or Gini index. Therefore, the time complexity of the train function is *O(d \* n \* 2^p).*

Time complexity of the predict function: *O(n \* p)*

For every sample predicted, the predict function would need to start from the root node and reach one of the leaf nodes, and compare the sample with every node traveled. At each level of the tree, one comparison is performed. If we assume the predicting size is also *n*, and since the maximum depth is *p*, the time complexity is *O(n \* p).*

Q2

表格

中度可信度描述已自动生成d)

From the table, we can see that all model assessment strategies, except the Holdout, yield Train AUC very close to the True Test’s Train AUC. As for ValAUC, both k-fold and Monte Carlo cross validation show larger ValAUC than Holdout does. However, the True Test’s ValAUC is larger than any other strategies. One thing noticeable here is that regardless of k-fold or Monte Carlo cross validation, increase of the k or s does necessarily result in the increase or decrease of the ValAUC. However, it does result in the increase of the Time taken to perform the assessment. When k and s is equal, Monte Carlo cross validation has a shorter assessment time than the k-fold cross validation does. In addition, the Holdout method has the minimum assessment time.

Q3

1. I used 5-fold cross validation, since 5-fold yields the highest ValAUC in the Q2.

文本

描述已自动生成Result of K-NN with different *K*:

For K-NN: the optimal *k* is *k=14*.

文本

低可信度描述已自动生成图片包含 日历

描述已自动生成Result of Decision Tree:

Figure 2 Different Min\_Sample\_Leaf with no Max\_Depth

Figure 1 Different Max\_Depth with Fixed Min\_Sample\_Leaf (1)

For Decision Tree:

The optimal Max\_Depth is *Max\_Depth=3*, and the optimal Min\_Sample\_Leaf is *Min\_Sample\_Leaf=50.*

文本

描述已自动生成d)

From the table, we can see that both K-NN and Decision Tree are not very sensitive to the training datasets. As more data is removed from the training dataset, there is no big impact on both algorithms’ accuracies and AUCs.