Submitted 3/10/2024

Decision tree classification and regression

Christopher Lee clee349@jhu.edu

Johns Hopkins University

Abstract

The decision tree is a popular learning method for regression and classification, because of its interpretability. The decision tree relies on the bias that the response variable can be predicted by a hierarchical sequence of linear discriminant decisions. As with all learning methods, bias and variance must be balanced in decision trees. Pruning provides a means of reducing the effects of overfitting, which decision trees are susceptible to when the training set is large. We expect that pruning may yield greater improvements on datasets were larger sample sizes or larger numbers of features, which may be more prone to overfitting. Full and post-pruned decision trees are constructed for six datasets. Decision tree classification and regression outperform a null model classifier and regressor except on the forest fires data set. Pruning was found to remove nodes in all data sets, but more strikingly for the regression data sets. Pruned trees also showed a larger improvement in performance for abalone and the forest fires data. The abalone data is also the largest data set in sample size, and features are measured to the finest precision, consistent with the hypothesis that the full decision tree overfits this dataset.

Problem statement

Decision trees offer a non-parametric approach to classification and regression problems. This approach solves prediction by separating feature sets into a sequence of linear discrimination decisions. When each individual decision is univariate, the decision tree is easier to interpret. The interpretability of decision trees can be one of its greatest strengths.

Like all learning algorithms, decision trees have inductive biases and the ability of learned decision trees to generalize to new data must be balanced with the degree of variance. As the size of the training set increases, decision trees will learn more decisions to predict the label, and the learned decisions are more likely to follow variance that is specific to the training set and does not generalize to unseen data. One method to counteract the risk of overfitting data is pruning the decision trees, which can be implemented as the tree is being built (pre-pruning), or after a full tree has been built (post-pruning). Specifically, tree nodes are pruned (child nodes are collapsed to the parent node) if doing so improves performance on data that was held out from training.

In this study, we compare the performance of full decision tree classification and regression with that of decision tree prediction with post-pruning. Because pruning may remove decisions that overfit the training data, we expect that a pruned decision tree will outperform the full decision tree if the full decision tree is overfitting the data. A decision tree will more likely overfit the data if there are more observations, more features, or

more distinct levels of each feature, as each "patch" of feature-space populated by training examples will be memorized to a value or class.

Experimental approach

To implement decision tree classification and regression, a generic tree object is defined in Python class 'DecisionTree'. The object contains a few fields, including the 'children' field for pointers to the child nodes of the object. The method 'predict' is also defined in the class.

A decision tree is generated from the training data through a recursive implementation. The method 'Generate_Tree' checks if all training data contained in the node are identical across features. If all data contain the same feature values, then the data cannot be split in the feature space, and the node is assigned the majority label value (classification) or mean label value (regression), and is assigned no child nodes (it is a leaf node). The method will also designate a node as a leaf node if its measure of purity (entropy for classification and variance for regression) exceeds a threshold (falls below the threshold in absolute value), allowing for pre-pruning. In this study, we build the decision trees fully, with no pre-pruning, by setting the threshold to zero. However, the threshold parameter allows for pre-pruned trees to be constructed.

If a node is not determined to be a child node, the feature that produces the best splits is identified. Here, the best splits are defined by maximizing gain ratio (classification), or minimizing mean sum squared error (regression). Additionally, for numeric features, splits at each mid-point between unique sorted values are compared. When the best feature and splitting criteria are determined, they are used to split the node data, and each segment is used to recursively generate the child nodes of the current node.

After the decision tree has been fully built, the tree is pruned by comparing the performance of the full tree to alternative versions of the tree where non-leaf nodes (parent nodes) are pruned. Post-pruning is implemented in method prune() and was the most challenging component to design. First, a list of all non-leaf nodes is constructed recursively. Here, each node is described with a location code, corresponding to the sequence of child nodes needed to navigate from the root to the node. For example, a node with location (0, 0, 1) is the root's 0^{th} child's 0^{th} child's 1^{st} child. Then, we iterate through the non-leaf nodes, generating a copy of the full decision tree, but with the iterated node replaced by a leaf node. If the performance of this pruned tree is better than the full tree, prune() returns the pruned tree. However, to check remaining nodes, prune is repeated iteratively, until all non-leaf nodes have been checked and are unable to produce a pruned tree that with better performance than the base tree. This iterative process is implemented in iter_prune().

Results

Classification problems

Breast Cancer data

The full decision tree showed high accuracy in a Kx2 classification, with all five folds exceeding 90% accuracy, and averaging 94%. The pruned decision tree performance was very similar. Both decision trees perform much better than the null model, which has an accuracy of 67%.

Table 1. Decision Tree Performance and Size for Breast Cancer Data

Fold	Full tree accuracy	Pruned tree accuracy	Full tree parent nodes	Pruned tree parent nodes
0	0.953571	0.953571	202	202
1	0.960714	0.960714	203	203
2	0.925	0.942857	189	101
3	0.953571	0.953571	195	195
4	0.928571	0.95	196	188

The pruned tree was the same size as the full tree in 3 folds; that is, no nodes were pruned. In one fold, tree size was reduced considerably by pruning, from 189 to 101. The first feature split is Clump Thickness at 6.5

Car evaluation

The full decision tree showed somewhat lower accuracy compared to the breast cancer data, with 90% average accuracy. The pruned decision tree performance was very similar across all folds. Both decision trees perform above the null model, where accuracy was 69%.

Table 2. Decision Tree Performance and Size for Car Evaluation Data

Fold	Full tree accuracy	Pruned tree accuracy	Full tree parent nodes	Pruned tree parent nodes
0	0.894356	0.894356	417	409
1	0.885673	0.890014	408	406
2	0.89725	0.904486	401	391
3	0.908828	0.911722	401	397
4	0.913169	0.913169	405	398

In all five folds, few nodes were removed during pruning, and accuracy was slightly higher, or as high as the full tree.

House votes

The full decision tree showed somewhat high accuracy on house votes data, similar to performance on breast cancer data, with 93% average accuracy for the full tree and 94% average accuracy for the pruned tree. In comparison, null model accuracy was only 6o%.

Table 3. Decision Tree Performance and Size for House Votes Data

Fold	Full tree accuracy	Pruned tree accuracy	Full tree parent nodes	Pruned tree parent nodes
0	0.954023	0.954023	138	138
1	0.936782	0.936782	118	118
2	0.896552	0.936782	130	49
3	0.965517	0.95977	132	53
4	0.948276	0.948276	132	132

While no nodes were pruned in three folds, more than half of nodes in the full tree were pruned for the other two folds. The first feature split was physician fee freeze.

Regression problems

Abalone

Regression using the full decision tree slightly outperforms the null model, with a mean squared error of 3.01, compared to 3.20 for the null model. The pruned tree shows a stronger improvement, reducing the mean squared error to 2.50.

The full decision tree was the largest tree generated, with over 1600 nodes in each of the 5 folds. This also required more time to generate. Post pruning greatly reduced the number of nodes, to about 10% of the full tree size.

Table 4. Decision Tree Performance and Size for Abalone Data

	Full tree mse	Pruned tree mse	Full tree parent nodes	Pruned tree parent nodes
0	2.99051	2.462252	1640	274
1	2.945749	2.573211	1616	119
2	2.966096	2.470821	1637	171
3	3.109992	2.502923	1623	152
4	3.02503	2.477383	1629	138

Computer Hardware

Decision tree performance was considerably better than the null model for the computer hardware data. While null model performance has mean squared error of 136.5, the mse of the full decision tree is 96.6 and the pruned tree mse is 95.5. Although the error of the full and pruned trees are similar, the pruned tree does have much fewer nodes.

Table 5. Decision Tree Performance and Size for Computer Hardware Data

Fold	Full tree mse	Pruned tree mse	Full tree parent nodes	Pruned tree parent nodes
0	71.222259	70.605922	79	29
1	67.445679	65.138966	78	51
2	145.480955	145.448938	79	40
3	92.653797	88.675701	81	16
4	106.236988	107.715267	81	24

Forest Fires

Decision tree regression performance was not better than the null model for the forest fire data, which is noted for being difficult to regress. The mean squared error of the null model, 64.5, is instead lower than those of the full and pruned decision trees, 75.6 and 67.8, respectively. The size of the pruned trees is much smaller than the full tree. Only a few nodes, and in some cases one node make up the pruned trees, while the full trees contain over 200 nodes.

	Full tree mse	Pruned tree mse	Full tree parent nodes	Pruned tree parent nodes
0	40.376589	32.647233	203	1
1	101.99176	99.465407	203	2
2	86.128831	79.898614	201	7
3	103.096428	93.371979	206	1
4	99.485473	96.848742	203	5

Discussion

We find that a decision tree performs better than a null model on regression and classification, except for the forest fires data set. Classification accuracy is high on all three classification data sets.

More nodes were removed by pruning in the regression data than the classification data sets, and these changes were associated with more noticeable improvements in

performance. It's not clear if regression problems are generally more affected by pruning, or if it's coincidence that the three regression data sets were more affected by pruning. It should also be noted that two of the classification data sets contain categorical features, and the third set with numeric features have all numeric measures on integer scales, reducing the number of distinct feature levels. In contrast, for the regression data, two of the data sets use decimal precision, while the third (computer hardware) uses integer precision.

Growing a full decision tree can hinder performance if the learning is overfitting the data, suggesting that improvements from pruning indicates that the full decision tree may be overfitting. The largest improvements from pruning were seen with the abalone and forest fires data sets. The abalone data set was the largest, and it is reasonable to expect that generating a full decision tree to match all of the feature variations will lead to overfitting. The forest fires data set is known to be difficult to regress, and the pruned trees contained only a handful of nodes. It seems like the regression of the forest fires has a low signal to noise ratio. That is, the area of forest fires has a dependence on only a few decisions, and those dependencies may be rather weak.

Conclusion

The decision tree provides good classification accuracy and moderate regression error, similar to k-nearest neighbors classification and k-means regression. Pruning most strongly improved performance and reduced tree size for abalone and forest fires data sets, which may be overfit when the decision tree is fully generated. The full decision tree for abalone likely overfits the data, as all features are measured with higher precision, and the dataset size is large.