Intermediate Data Science

Decision Trees

Joanna Bieri DATA201

Mini Review - Decision Trees

REVIEW

- Decision Trees use a series of splits to recursively split feature space into regions that divide up target variables.
- They are interpretable, meaning we can see how they are making their decisions, and they do a good job of capturing nonlinear relationships.
- At each step the algorithm chooses the split that **reduces variance** (for regression) or **minimizes impurity** (for classification).

Mini Review - Gini Impurity and Entropy

REVIEW

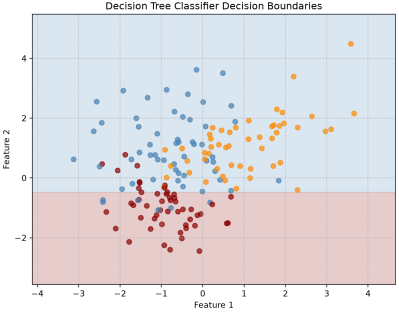
$$G = 1 - \sum p_k^2 \qquad H = -\sum_k p_k \log_2(p_k)$$

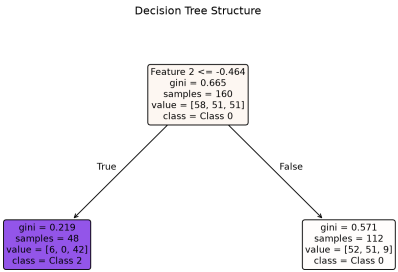
Situation	Gini	Entropy	Purity
100% one class	0.0	0.0	Pure
80/20 mix	0.32	0.72	Somewhat impure
50/50 mix	0.50	1.00	Very impure

At each node, the Decision Tree tries every possible split and picks the one that results in the biggest drop in impurity.

Mini Review - Decision Tree for Classification

REVIEW



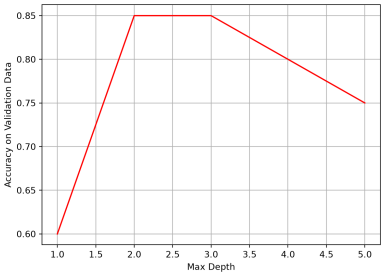



```
clf = DecisionTreeClassifier(max_depth=1, criterion='gini')
clf.fit(X_train, y_train)
```

Decision tree boundaries are piecewise-rectangular.

Mini Review - What depth should I use?

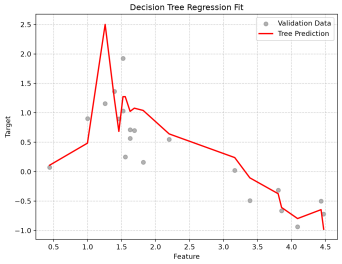
REVIEW



A for loop over `max_depth` $= 1$ to 5, accuracy on the **validation** data (train / validation / test, 80 / 10 / 10 like Day 13). Then the test set, once.

Mini Review - No pruning and no max-depth

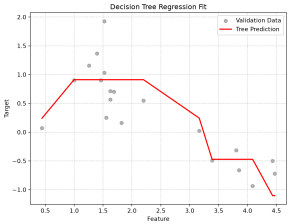
REVIEW



We had ZERO error on our training data and an R2 of 1!!! We are memorizing the training data.

Mini Review - Controlling Tree Complexity

REVIEW

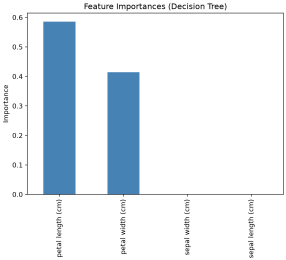



```
tree_reg = DecisionTreeRegressor(ccp_alpha=0.01)
```

Pruning removes branches that provide little improvement in predictive power.

Mini Review - Feature Importance

REVIEW



`tree.feature_importances_`

Higher values mean the feature contributed more to reducing impurity.

Impurity

ASK

Node A: 10 dogs and 10 cats.

Node B: 18 dogs and 2 cats.

Which node is more pure? What is the Gini of each?

Hint. $G = 1 - \sum p_k^2$, and the proportions in A are 0.5 and 0.5.

Impurity

ANSWER

Node A: 10 dogs, 10 cats. Node B: 18 dogs, 2 cats. Which is more pure? Gini of each?

```
A: 1 - (0.5^2 + 0.5^2) = 1 - 0.50 = 0.50
B: 1 - (0.9^2 + 0.1^2) = 1 - 0.82 = 0.18
```

Node B is more pure, its Gini is closer to 0. A 50/50 node is as impure as two classes can get.

Impurity

ASK

Why do the boundaries of a decision tree always look like rectangles?

Hint. Think about what one question in the tree looks like.

Impurity

ANSWER

Why do the boundaries of a decision tree always look like rectangles?

- Every question in the tree is about ONE feature: is feature 1 $\leq$ some number?
- That is a straight line across that one axis.
- Stack a few of those questions and you get boxes.

Do Trees Need Scaling?

ASK

On Day 13 we said ALWAYS SCALE.

Does a decision tree need

StandardScaler? Why or why not?

Hint. KNN needed scaling because it measures distance. What does a tree measure?

Do Trees Need Scaling?

ANSWER

Does a decision tree need StandardScaler? Why or why not?

- No. A tree only asks "is this feature $\leq$ some cutoff?", one feature at a time.
- If you rescale the feature, the cutoff rescales too and the split is the same.
- It does not hurt, it just does not help. KNN and regularization still need it.

Overfitting

ASK

A tree with no max_depth and no pruning:
training MSE is 0, validation MSE is much bigger.
What happened? What would you change?

Hint. How many training points can end up in one leaf?

Overfitting

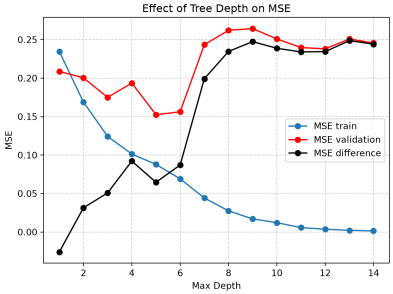
ANSWER

No max_depth, no pruning: training MSE 0, validation MSE much bigger. What happened?

- The tree kept splitting until every leaf had one training point. It memorized the training data.
- That is overfitting, the same thing we saw with $k = 1$ in KNN.
- Limit `max_depth`, or prune with `ccp_alpha`.

Overfitting

ASK



Which max depth would you pick? Why not the depth with the smallest difference?

Overfitting

ANSWER

Which max depth would you pick? Why not the depth with the smallest difference?

- Pick the depth where the validation MSE is lowest, here 5. Then test that one tree, once.
- Depth 1 has the smallest difference, but both errors are high. That tree is too simple, it underfits.
- Past depth 6 the training MSE keeps falling and the validation MSE goes UP.

Overfitting

ASK

What happens to the tree

as you make ccp_alpha bigger?

What happens to the training error?

Hint. Remember alpha in Ridge and Lasso: a penalty for a complicated model.

Overfitting

ANSWER

What happens to the tree as you make ccp_alpha bigger? To the training error?

- Bigger alpha cuts off more branches, so the tree gets smaller (fewer leaves).
- The training error goes up, on purpose. With a good alpha the validation error goes down.
- Same idea as Ridge and Lasso: a penalty for complexity.

Feature Importance

ASK

A feature gets importance 0 in your tree.
Does that mean it has nothing to do
with the target?

Hint. In the Iris data, petal length and petal width go up together.

Feature Importance

ANSWER

A feature gets importance 0. Does that mean it has nothing to do with the target?

- Not always. It means THIS tree did not use it to split.
- If two features carry the same information, the tree uses one and the other gets 0.
- A depth limit can also leave a useful feature out.

INTERVIEW

How does a decision tree decide where to split?

INTERVIEW

Why do decision trees overfit,

and how do you stop it?

Questions?