DA2011
Machine Learning I

Lecture 4

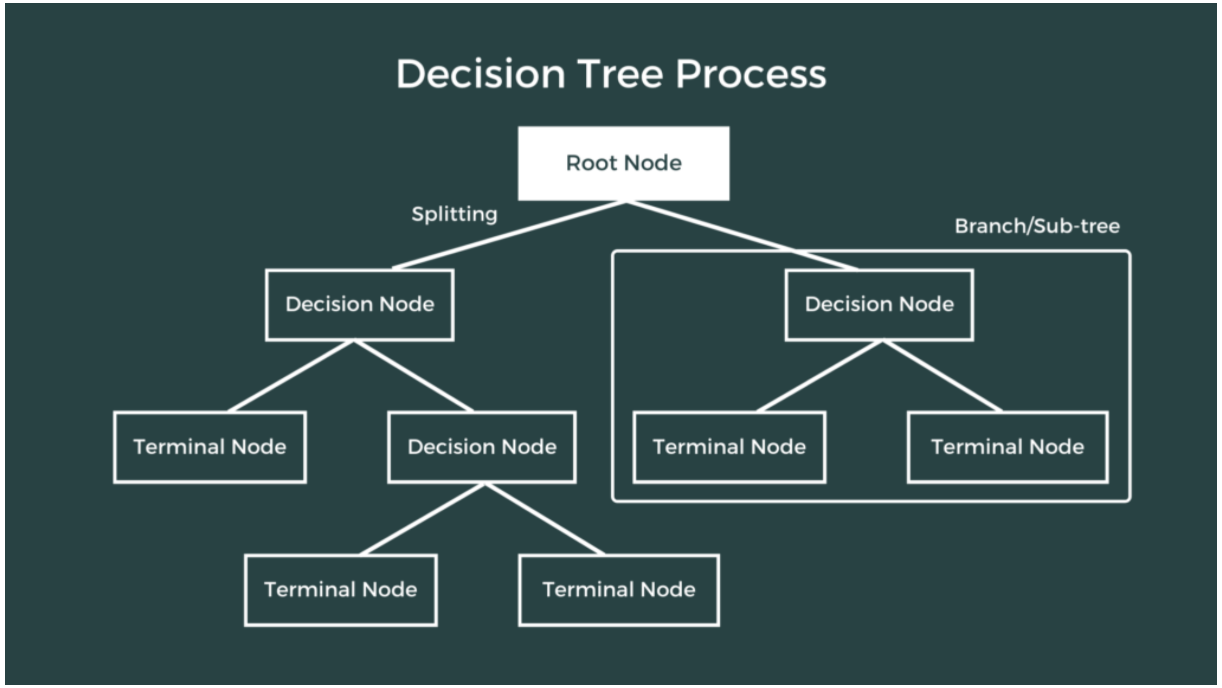
Ms. Naethree Premnath
12th October 2025

Today you will learn…

- Decision Tree
  - –Decision Tree for Regression
  - –Feature Selection for Regression Trees
  - –Decision Tree for Regression Example
  - –Decision Tree for Classification
  - –Feature Selection for Regression Trees
  - –Parameters in Decision Trees
  - –Overfitting in Decision Trees
  - –Pruning in Decision Trees

Decision Trees

- A Decision Tree is a supervised learning algorithm used for both regression and classification.
- It models decisions as a tree-like structure.



- Each internal node → a condition on a feature (e.g., "Age > 30?")
- Each branch → an outcome of the condition (Yes/No)
- Each leaf node → a prediction (numerical value for regression or class label for classification)

It works recursively such that the dataset is split into smaller subsets until a stopping criterion is met.

Decision Trees

A Decision Tree builds itself top-down using a process called **Recursive Binary Splitting**.

- **Recursive**: The process repeats at each node, once the data is split, the same steps are applied to each subset.
- **Binary**: Each split divides the data into two groups based on one feature and a threshold value.
- **Splitting Criterion**: The algorithm chooses the best feature and threshold that most effectively separates the data.

The Goal : Reduce the prediction error (make each region as “pure” (homogeneous) as possible in terms of target values)

Feature Selection for Regression Trees

In regression, the goal is to predict a continuous value. We want to create nodes where the target values are as close to each other as possible (low variance).

Key Metric:

- Variance Reduction / MSE

$$\text{Variance Reduction / MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

Regression Tree Example

Let's consider an example of **predicting house price** using variables 'area' and 'location'

Example data:

	Predictors		Response
House	**Area(sqft)**	**Location**	**Price($1000s)**
1	1200	Urban	240
2	1500	Urban	260
3	1700	Rural	200
4	2000	Urban	300
5	2300	Rural	220
6	2600	Urban	340

Regression Tree Example

Step 1: Calculate the Variance before splitting

House	Area(sqft)	Location	Price($1000s)
1	1200	Urban	240
2	1500	Urban	260
3	1700	Rural	200
4	2000	Urban	300
5	2300	Rural	220
6	2600	Urban	340

$$Mean = \frac{240 + 260 + 200 + 300 + 220 + 340}{6} = 260$$

$$Variance = \frac{(240-260)^2 + (260-260)^2 + (200-260)^2 + (300-260)^2 + (220-260)^2 + (340-260)^2}{6} = 2266.67$$

Regression Tree Example

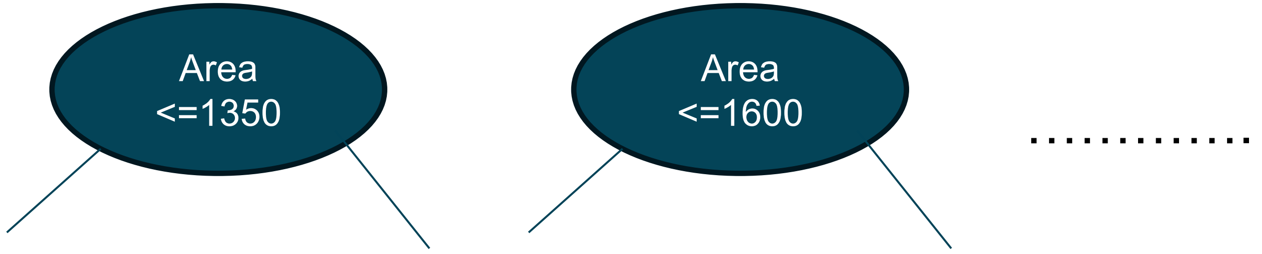
Step 2: Define a split
At the root node, the algorithm will try every possible split across both features (Area and Location).

Option 1: Split by Area

For the numerical feature Area, the algorithm will test possible thresholds (midpoints between sorted values):

Sorted Area values → 1200, 1500, 1700, 2000, 2300, 2600
Possible split points → 1350, 1600, 1850, 2150, 2450

Example of testing splits:



Regression Tree Example

Let's consider just one of the split

Area
<=1850

House	Area(sqft)	Location	Price($1000s)
1	1200	Urban	240
2	1500	Urban	260
3	1700	Rural	200

Mean price = 233.3
Variance = 622.22

House	Area(sqft)	Location	Price($1000s)
4	2000	Urban	300
5	2300	Rural	220
6	2600	Urban	340

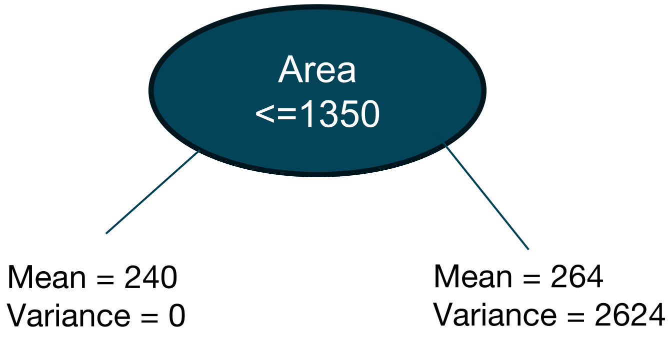
Mean price = 286.67
Variance = 2488.89

$$\boldsymbol{Weighted\ Variance = \frac{3}{6}(622.22) + \frac{3}{6}(2488.89) = 1555.56}$$

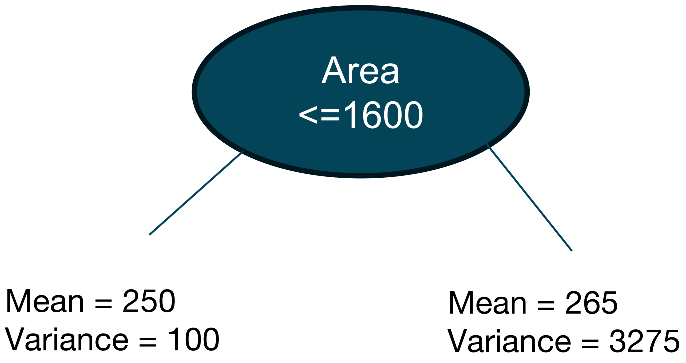
$$\boldsymbol{Variance\ Reduction = 2266.67 - 1555.56 = 711.11}$$

Regression Tree Example

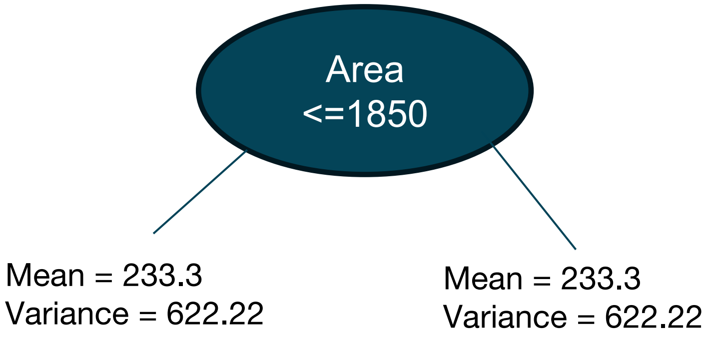
Let's consider each of the splits



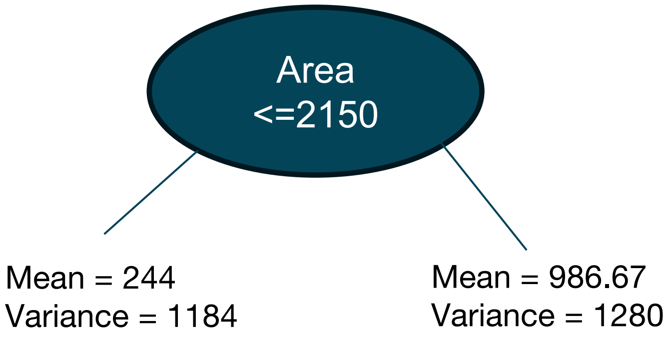
Weighted Variance = 2186.67
Variance reduction = 80.00



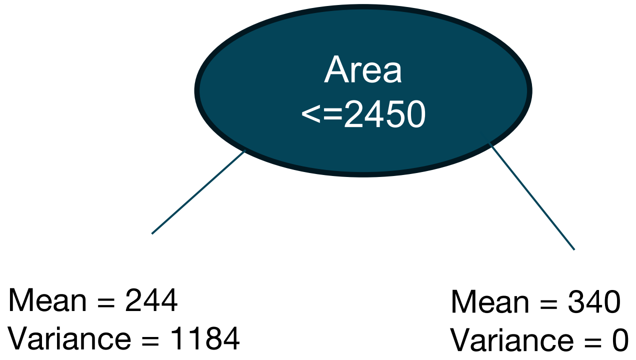
Weighted Variance = 2216.67
Variance reduction = 50.00



Weighted Variance = 1555.56
Variance reduction = 2266.67-1555.56 = 711.11



Weighted Variance = 986.67
Variance reduction = 1280



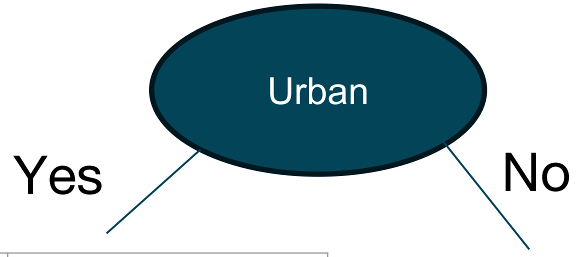
Weighted Variance = 986.67
Variance reduction = 1280

Regression Tree Example

Option 2: Split by Location

For the categorical feature Area, the algorithm will test possible values (Eg: Urban, Rural)

Example of testing the split:



House	Area(sqft)	Location	Price($1000s)
1	1200	Urban	240
2	1500	Urban	260
4	2000	Urban	300
6	2600	Urban	340

Mean price = 285
Variance = 1475

House	Area(sqft)	Location	Price($1000s)
3	1700	Rural	200
5	2300	Rural	220

Mean price = 210
Variance = 100

Weighted Variance after split = 1016.67
Variance reduction = 2266.67 – 1016.67 = 1250.00

Regression Tree Example

Step 3: Compare the Variance Reduction and choose the one with the highest reduction

Split by Area (<= 2150) and Split by Area (<= 2450) gives the largest reduction

How to choose among the two?

- Typically choose the smaller threshold (2150) as it's more general

House	Area(sqft)	Location	Price($1000s)
1	1200	Urban	240
2	1500	Urban	260
3	1700	Rural	200
4	2000	Urban	300

House	Area(sqft)	Location	Price($1000s)
5	2300	Rural	220
6	2600	Urban	340

Regression Tree Example

Step 4: Continue Splitting – Consider Right Node
Option 1: Split by Location

Area
<=2150

Yes

House	Area(sqft)	Location	Price($1000s)
1	1200	Urban	240
2	1500	Urban	260
3	1700	Rural	200
4	2000	Urban	300

No

House	Area(sqft)	Location	Price($1000s)
5	2300	Rural	220
6	2600	Urban	340

Urban

Yes

Mean price = 340
Variance = 0

No

Mean price = 220
Variance = 0

Before splitting, Variance = 3600
Weighted Variance after split = 0
Variance reduction = 3600-0 = 3600

Regression Tree Example

Option 2: Split by Area

House	Area(sqft)	Location	Price($1000s)
1	1200	Urban	240
2	1500	Urban	260
3	1700	Rural	200
4	2000	Urban	300

House	Area(sqft)	Location	Price($1000s)
5	2300	Rural	220
6	2600	Urban	340

Sorted Area values → [2300,2600]
Possible split point → 2450

Weighted Variance after split = 0
Variance reduction = 3600-0 = 3600

Regression Tree Example

Step 5: Choose the one with highest Variance Reduction

In this case, since both Location and Area both split to pure nodes you can choose either. Let’s choose Location since it is simpler.

Area
<=2150

Yes

House	Area(sqft)	Location	Price($1000s)
1	1200	Urban	240
2	1500	Urban	260
3	1700	Rural	200
4	2000	Urban	300

No

House	Area(sqft)	Location	Price($1000s)
5	2300	Rural	220
6	2600	Urban	340

Urban

Yes

Mean price = 340
Variance = 0

No

Mean price = 220
Variance = 0

Weighted Variance after split = 0
Variance reduction = 3600-0 = 3600

Take Home Activity

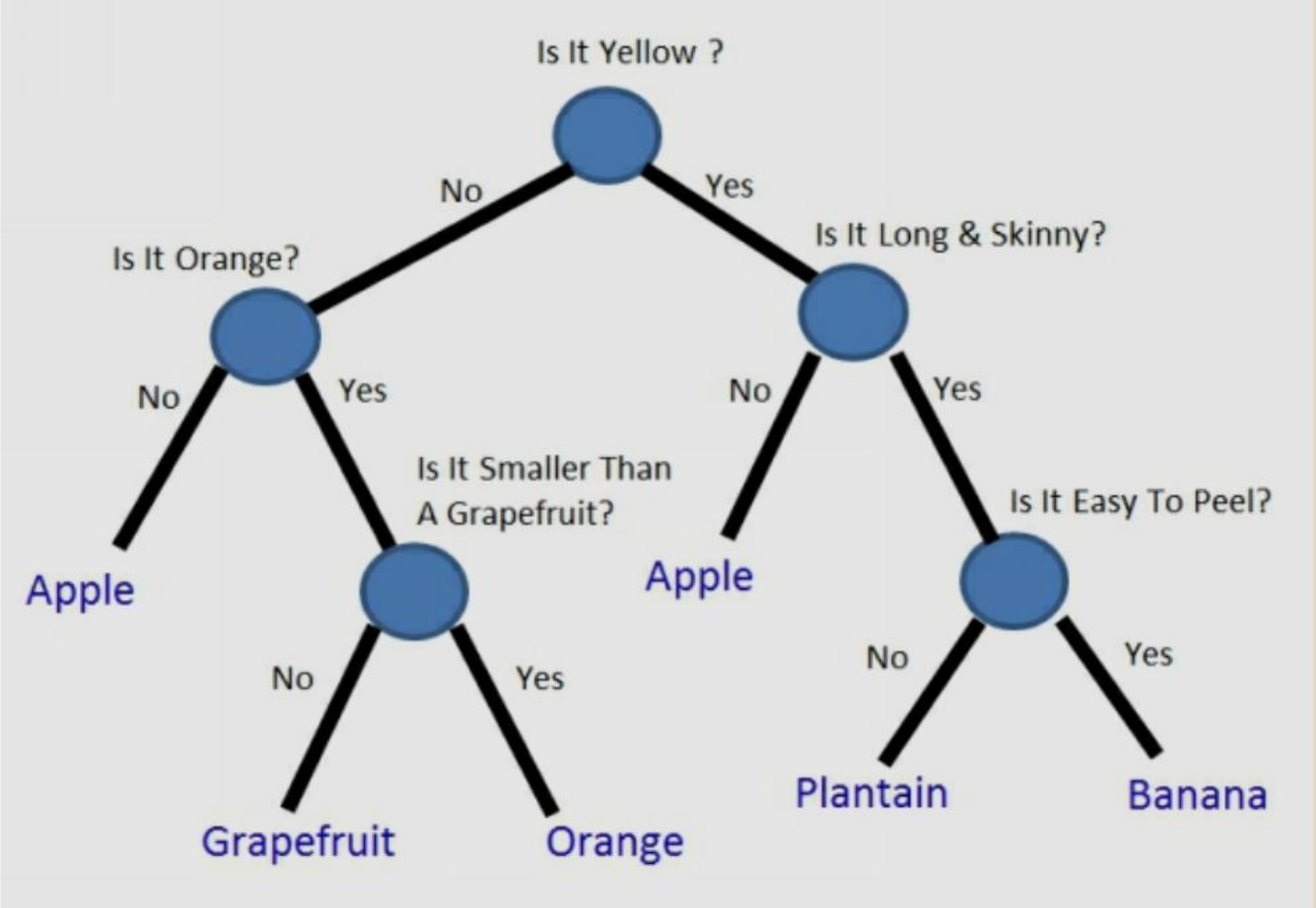
The right node is already complete. Continue the process for the left node until all nodes become pure.

- Make a copy of the document in this sheet and fill it. (To fill it, you can double click the diagram and directly edit it)

Decision Tree for Classification

Predict categorical outcomes (e.g., Yes/No, A/B/C).

E.g:



Feature Selection For Classification Trees

In classification, the goal is to create nodes where the majority of data points belong to a single class.

Key Metrics:

1. Gini Index

$$Gini = 1 - \sum_{i=1}^{C} {p_i}^2$$

2. Entropy (Information Gain)

$$Entropy = -\sum_{i=1}^{C} p_i \log_2 p_i$$

p_i : $proportion\ of\ samples\ in\ the\ node\ that\ belong\ to\ class\ i$

C: $total\ no.\ of\ classes\ in\ the\ target$

i: $index\ for\ the\ classes$

Decision Tree for Classification

Step 1: Calculate Gini or Entropy at root.

Step 2: Consider possible splits.

Step 3: Calculate weighted impurity for each split.

Step 4: Choose split with lowest impurity.

Parameters in Decision Trees

- **Max Depth** - The maximum depth of the tree. Limits how deep the tree can grow.
- **Min Samples Split** - The minimum number of samples required in a node before it can be split.
- **Min Samples Leaf** - The minimum number of samples that must be in a leaf node.
- **Min Impurity Decrease** - A node will only split if the reduction in impurity (like Gini or MSE) is greater than this value.
- **Max Features** - Number of features to consider when looking for the best split.
- **Max Leaf Nodes** - The maximum number of leaf nodes allowed in the tree.

Overfitting in Decision Trees

The Overfit Tree:

- Very deep and complex.
- Many branches for minor patterns.
- Perfect accuracy on training data, poor accuracy on test data.

Our Goal: Find a tree that is accurate but also simple and generalizable.

Pruning in Decision Trees

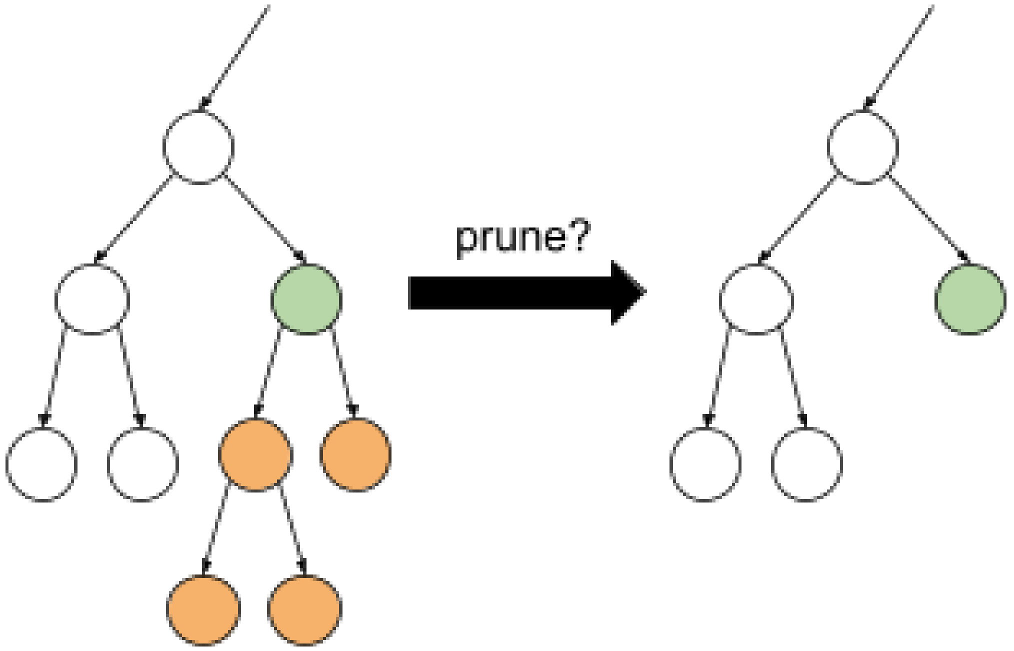
Why Pruning?

- Prevents overfitting
- Reduces model complexity
- Improves generalization
- Enhances interpretability
- Facilitates Model Maintenance

Types of Pruning:

1. Pre-Pruning (Early Stopping)

2. Post-Pruning (Prune after tree is grown)



Pre-Pruning

Common pre-pruning techniques:

- **Maximum Depth**: It limits the maximum level of depth in a decision tree.
- **Minimum Samples per Leaf**: Set a minimum threshold for the number of samples in each leaf node.
- **Minimum Samples per Split**: Specify the minimal number of samples needed to break up a node.
- **Maximum Features**: Restrict the quantity of features considered for splitting.

By pruning early, we come to be with a simpler tree that is less likely to overfit the training facts.

Post-Pruning

Common post-pruning techniques include:

- **Reduced Error Pruning**: Removes branches that do not significantly affect the overall accuracy.

- **Minimum Impurity Decrease**: Prunes nodes if the decrease in impurity (Gini impurity or entropy) is beneath a certain threshold.

- **Minimum Leaf Size**: Removes leaf nodes with fewer samples than a specified threshold.

Thank you
See you next week!