Decision Trees and Different Metrics

TM Quest

Overview

What Will we Learn in This Module?

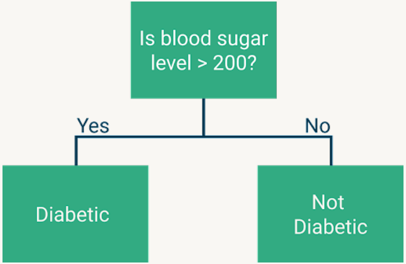
- What are decision trees?
- How to use decision trees for regression/classification
- How to visualize decision trees
- How to tackle unbalanced datasets
- What is precision and recall?
- What is the precision-recall trade-off?

Introduction to Decision Trees

What are Decision Trees?

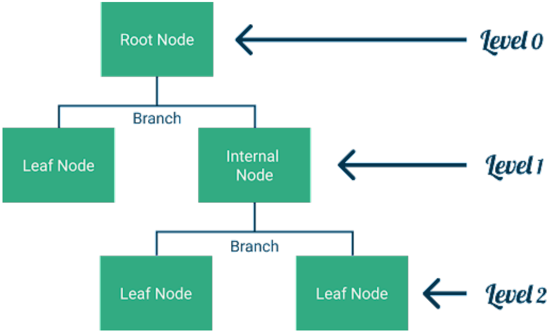
Example

- Feature: The blood sugar level of a patient.
- Target: Does the patient have diabetes.
- Simple Decision Tree:
  - Blood sugar > 200 $\mathrm{mg/dl} \implies$ is diabetic
  - Blood sugar ≤ 200 $\mathrm{mg/dl} \implies$ is not diabetic



Some Tree Terminology

- Root of a tree is where the tree starts.
- Branching point is where the tree splits.
- The end points of a tree are called leafs or terminal nodes.
- All nodes which are not leafs are called internal nodes.
- The level of a node is how many steps one needs to take to go from the node to the root.
- The tree-depth is the maximal level of the tree.



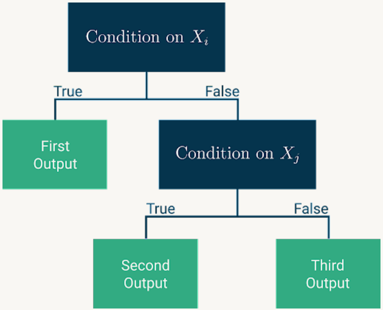
Decision Trees

Decision Trees

Definition

A decision tree model is a tree where:

- on each of the branching points we decide on going left or right based on a condition for one of the features.
- when a leaf node is reached the model outputs a predicted value.



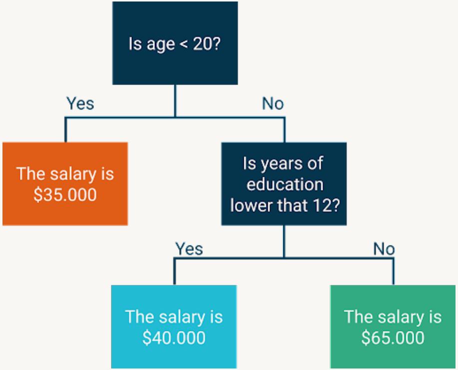
Training

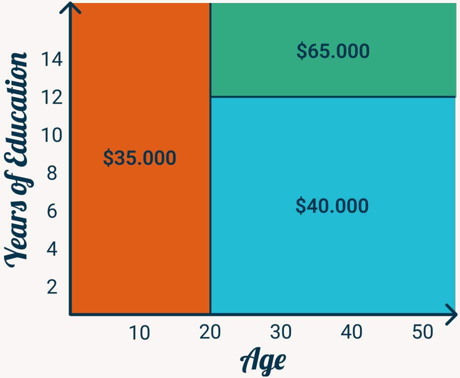
- The training step tries to find the best tree for the given training data.
- How this is done depends if we want to do a regression or a classification task.

Decision Tree Regression Example

Example

- Features: The age and years of education of a person.
- Target: The salary of the person.



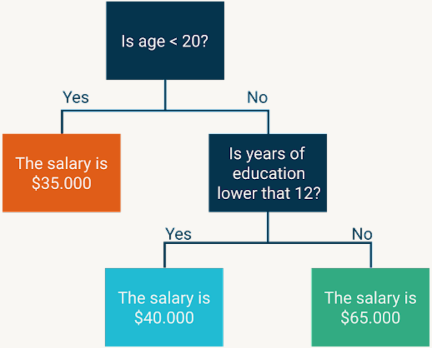


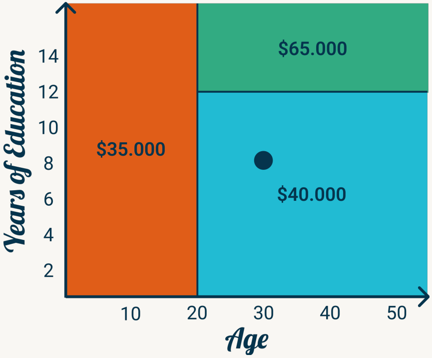
Decision Tree Regression Example

Example

Let us say that John

- has 8 years of education
- is 30 years old.





False Positive and False Negative

Unbalanced Datasets

Definition

We say that a dataset is unbalanced if the number of targets in the dataset in each category is very unequal in size.

Example

Let us say that we have a dataset predicting breast cancer from mammography.

- Feature: Mammography pictures.
- Target: Breast cancer Yes/No.

Let us say that the dataset contains the following targets:

- 295 did not have breast cancer
- 5 did have breast cancer.

This dataset is unbalanced.

The Problem with Accuracy Score

Accuracy Score Reminder

$$\frac{\text{Number of Correctly Classified Observations}}{\text{Total Observations}}.$$

Example

If the training data contains the following targets:

- 295 did not have breast cancer
- 5 did have breast cancer.

Then the model saying nobody have breast cancer have a 98% accuracy score!

False Positives and False Negatives

In binary classification we give out either the values True or False.

Two Types of Errors

- False positive is when the model predict true, while the actual value is false.
- False negative is when the model predict false, while the actual value is true.
- True positive is when both values are true.
- True negative is when both values are false.

		Actual Class	
		True	False
Predicted Class	True	True Positives	False Positives
	False	False Negatives	True Negatives

Precision and Recall

Precision and Recall

Definition (Precision and Recall)

$$\text{Precision} = \frac{\text{Number of True Positives}}{\text{Total Number of Predicted Positives}}$$

$$\text{Recall} = \frac{\text{Number of True Positives}}{\text{Total Number of Actual True Values}}$$

		Actual Class	
		True	False
Predicted Class	True	True Positives	False Positives
	False	False Negatives	True Negatives

Intuition

- High precision implies few false positives.
- High recall implies few false negatives.

Example of Precision and Recall

Example

		Actual Class	
		True	False
Predicted Class	True	True Positives 80	False Positives 20
	False	False Negatives 10	True Negatives 100

$$\text{Precision} = \frac{80}{80 + 20} = 80\%$$

$$\text{Recall} = \frac{80}{80 + 10} = 89\%$$

Precision-Recall Tradeoff

Precision-Recall Tradeoff

- Weighting precision higher will make the recall drop, and vice versa.
- Depending on the application, we might want high recall or precision.

Example

Let us say that we have a dataset predicting breast cancer from mammography.

- 295 did not have breast cancer
- 5 did have breast cancer.

Then we want to have high recall (few false negatives) and might accept lower precision (more false positives).