EEP 596: Adv Intro ML || Lecture 6

Dr. Karthik Mohan

Univ. of Washington, Seattle

January 24, 2023

Logistics

1. **Lightning Presentation Slot:** Please pick a slot for your 5 minute lightning presentation this quarter if not done already. Spreadsheet available on discord
2. **Conceptual 2:** Assigned and due the coming Sunday
3. **Programming 3:** Will be assigned today and will be a mini-project based on Kaggle contest - Due in about 2 weeks on February 4 (Saturday)
4. Anything else?

Last class

classification

1. How Logistic Regression differs from Linear Regression?
2. Evaluation metrics for Binary Classification F1-Score or AUC
3. Pre-processing and Feature engineering for Spam Classification
4. Bag of words model
5. TF-IDF

Bag of words

1) Vocab: V – Set of all possible words

S1: I like this ice cream

V: I, like, chocolate, icecream, food, beach

Bag of words for S1:

I	like	chocolate	icecream	food	beach
1	1	0	1	0	0

x_{S1} → Feature vector

$$x_{S1} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

What if $|V| = 100k$

Dense Embeddings ≠ Learned from Bag of words

↳ Learn about embeddings when we get to NLP/Deep Learning topics

Non-Contextual

Contextual embeddings

→ Representation Learning (ICLR)

This restaurant is not bad! →

not bad

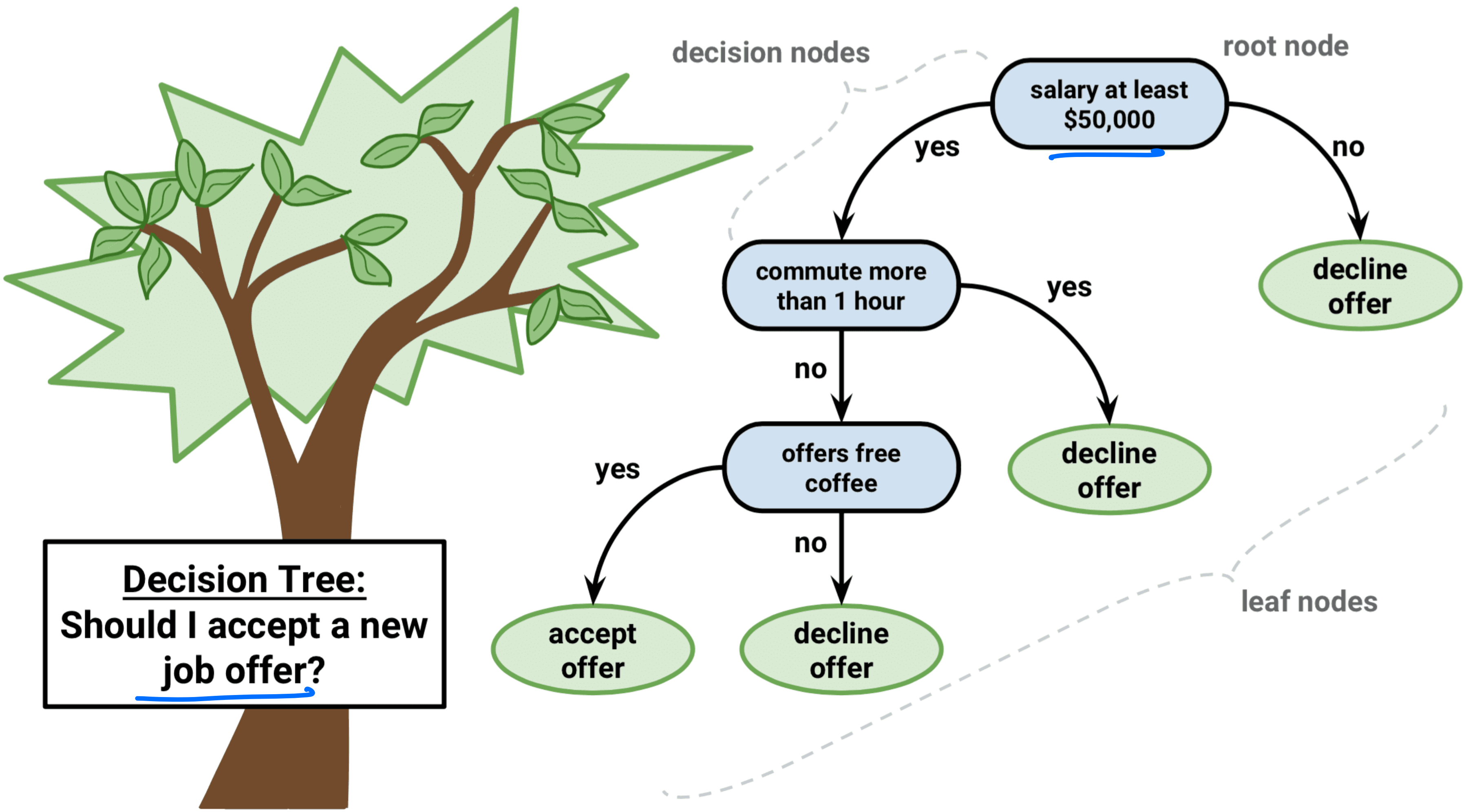
not_bad → Non-Linear feature

Today!

Decision Trees

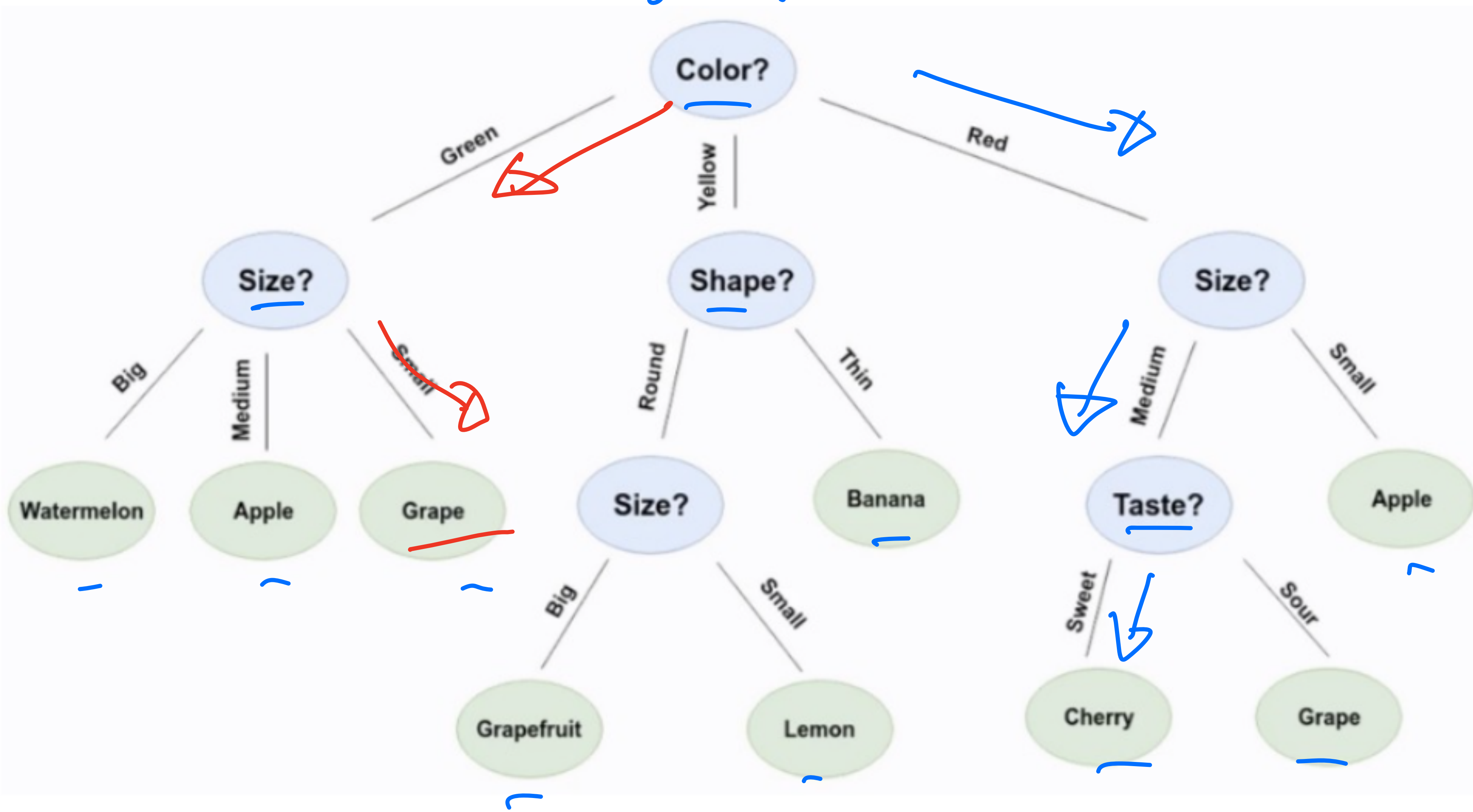
Next Topic: Decision Trees Classifier

Decision Trees Motivation



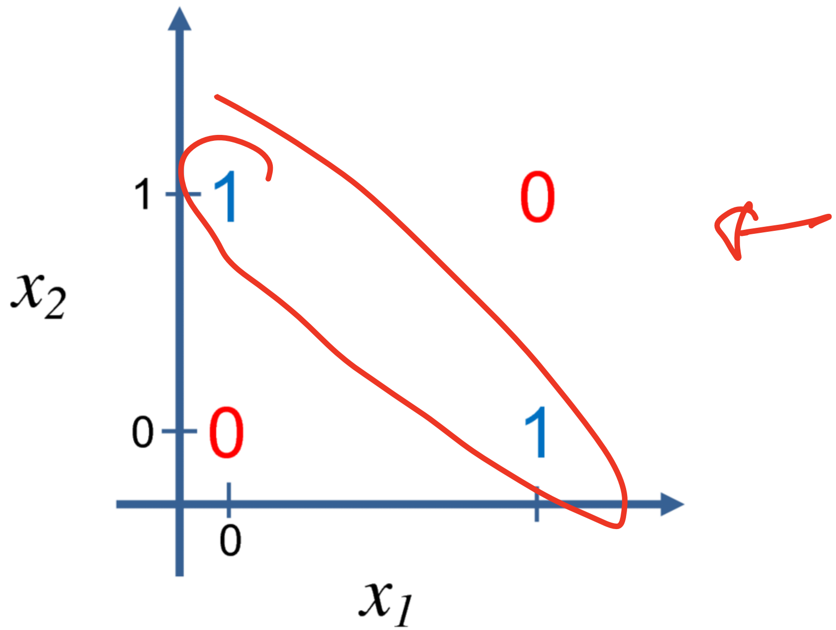
Decision Trees Motivation

Categorizing Fruits



ICE #1

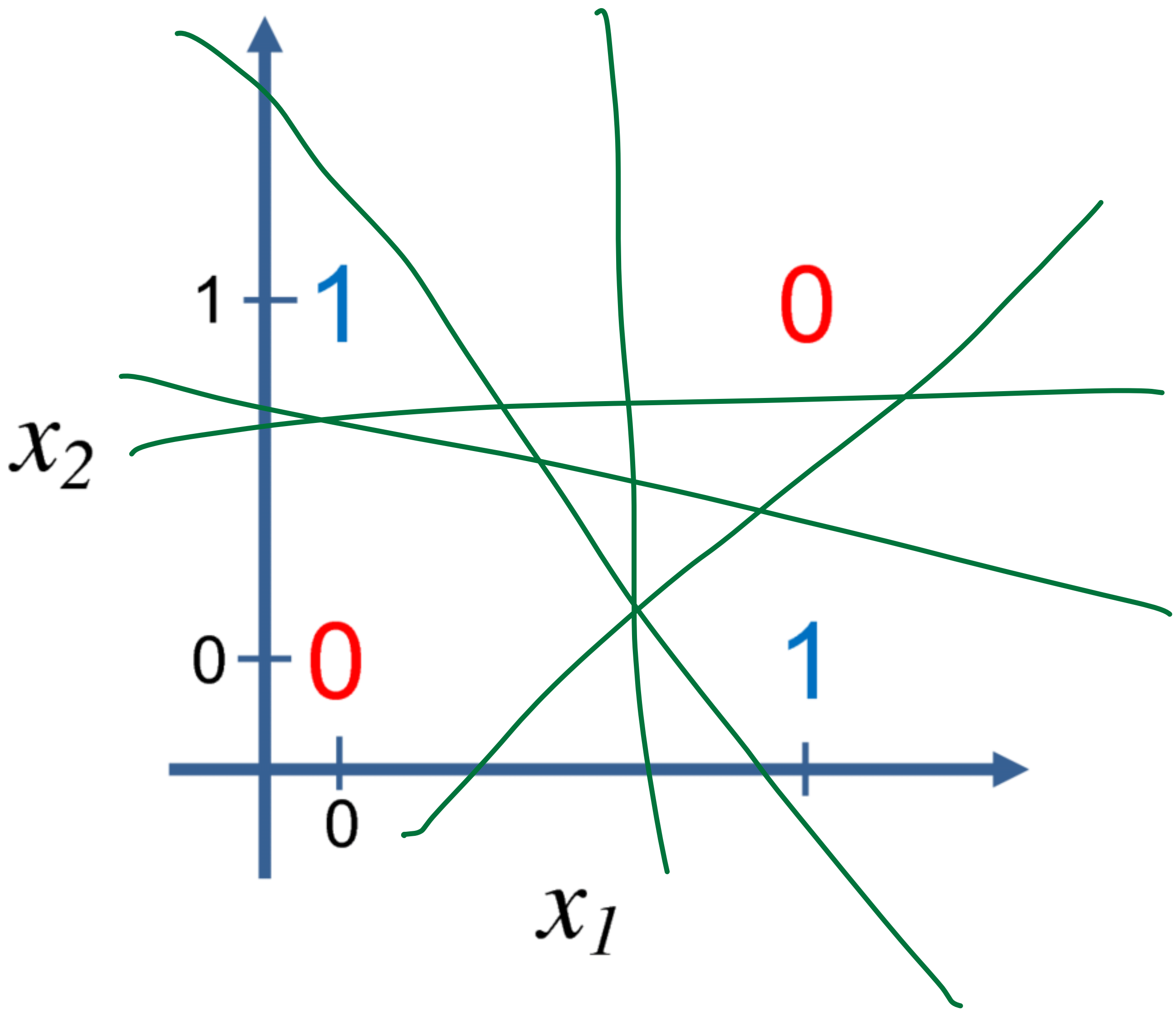
Can Logistic Regression learn to separate the 0's from the ones exactly?



1. Yes
2. No
3. Maybe

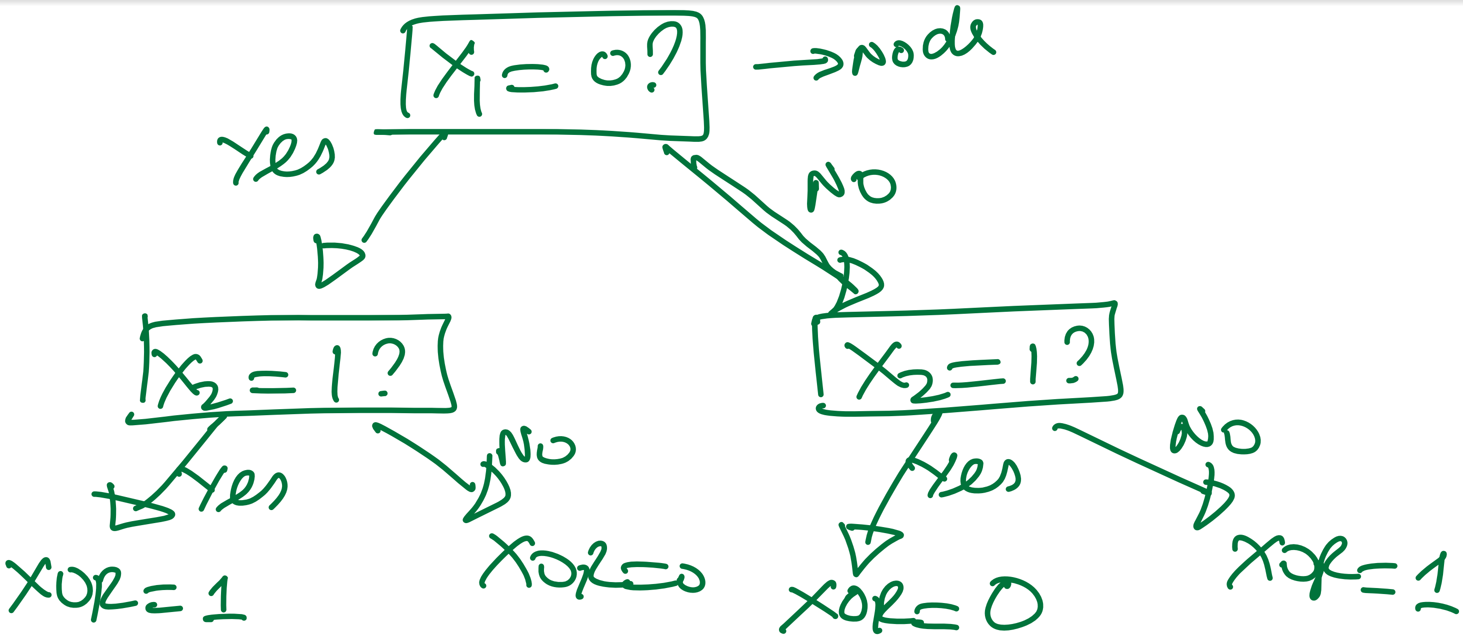
XOR Function

Linearly Separable?



XOR Function

Can XOR be modeled by Decision Tree?



Why Decision Trees?

1. **Human-like:** We usually make decisions based on if/then and else/or scnearios. **Example:** If it is raining outside, it's not too cold and it's summer time - Let's go hiking. **Example:** If it's raining and it's winter, let's skip hiking.

Why Decision Trees?

1. **Human-like:** We usually make decisions based on if/then and else/or scnearios. **Example:** If it is raining outside, it's not too cold and it's summer time - Let's go hiking. **Example:** If it's raining and it's winter, let's skip hiking.
2. **Explainability:** Medical AI is a good application area for Decision Trees. **Example:** Your *AI model for health care* predicts possible cancer from past health records and current CT scans. Both the patient and the doctor would like to know how the AI model arrived at this decision?

Why Decision Trees?

1. **Human-like:** We usually make decisions based on if/then and else/or scnearios. **Example:** If it is raining outside, it's not too cold and it's summer time - Let's go hiking. **Example:** If it's raining and it's winter, let's skip hiking.
2. **Explainability:** Medical AI is a good application area for Decision Trees. **Example:** Your *AI model for health care* predicts possible cancer from past health records and current CT scans. Both the patient and the doctor would like to know how the AI model arrived at this decision?
3. **Non-linear Models:** Decision Trees can easily model non-linear decision boundaries, unlike logistic regression. Modeling XOR or non-linear thinking in arriving at a decision.

Why Decision Trees?

1. **Human-like:** We usually make decisions based on if/then and else/or scnearios. **Example:** If it is raining outside, it's not too cold and it's summer time - Let's go hiking. **Example:** If it's raining and it's winter, let's skip hiking.
2. **Explainability:** Medical AI is a good application area for Decision Trees. **Example:** Your *AI model for health care* predicts possible cancer from past health records and current CT scans. Both the patient and the doctor would like to know how the AI model arrived at this decision?
3. **Non-linear Models:** Decision Trees can easily model non-linear decision boundaries, unlike logistic regression. Modeling XOR or non-linear thinking in arriving at a decision.
4. **Non-parametric:** Decision trees don't have the standard w parameter vector/weight vector. ω? $\times$

Why Decision Trees?

1. **Human-like:** We usually make decisions based on if/then and else/or scnearios. **Example:** If it is raining outside, it's not too cold and it's summer time - Let's go hiking. **Example:** If it's raining and it's winter, let's skip hiking.
2. **Explainability:** Medical AI is a good application area for Decision Trees. **Example:** Your *AI model for health care* predicts possible cancer from past health records and current CT scans. Both the patient and the doctor would like to know how the AI model arrived at this decision?
3. **Non-linear Models:** Decision Trees can easily model non-linear decision boundaries, unlike logistic regression. Modeling XOR or non-linear thinking in arriving at a decision.
4. **Non-parametric:** Decision trees don't have the standard w parameter vector/weight vector.
5. **Robustness to noise:** A few noisy examples in the data set may not through a decision tree prediction off - Based on majority votes.

Learning Decision Trees

Learning

The `learning` for Decision Trees boils down to how to build the tree. Which feature to split on first? Second? And so on... Also, when to stop building the tree

Learning Decision Trees

Learning

The `learning` for Decision Trees boils down to how to build the tree. Which feature to split on first? Second? And so on... Also, when to stop building the tree

Intuition behind building Decision Trees

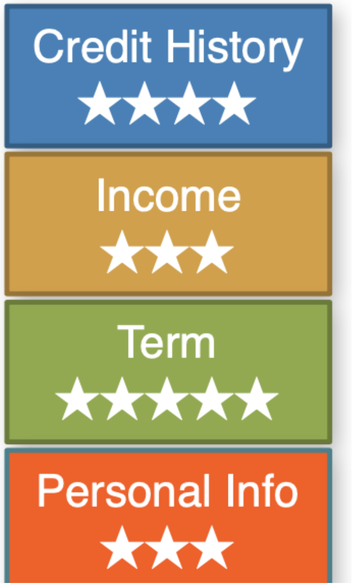
Start splitting on features that give the maximum information gain or reduce the uncertainty in prediction/reduce the classification error. This is done iteratively and hence can be thought of as a greedy procedure.

Case Study: What makes a loan risky?

Features: Credit History

Did I pay previous loans on time?

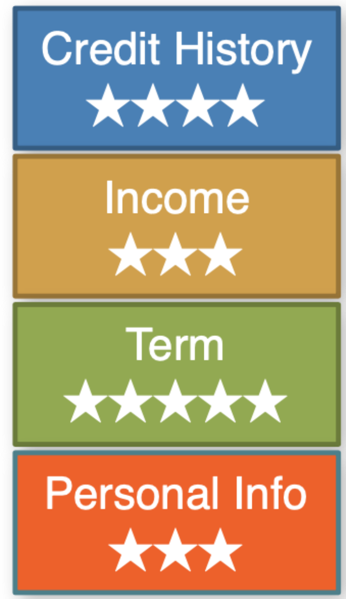
Example: excellent, good, or fair



Features: Income

What's my income?

Example:
$80K per year

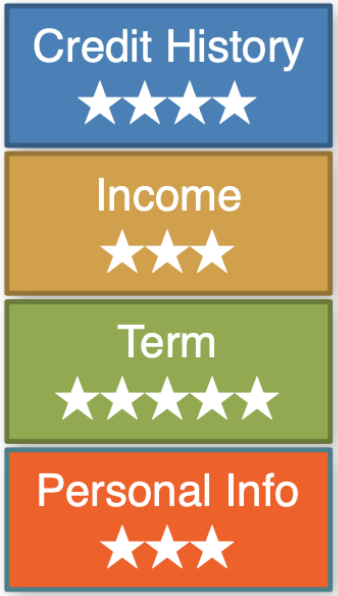


Features: Loan Terms

How soon do I need to pay the loan?

Example: 3 years, 5 years,…

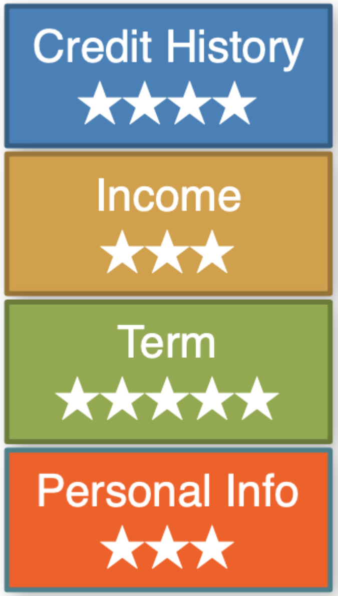
Home: - 15yrs or 30yrs



Features: Personal Information

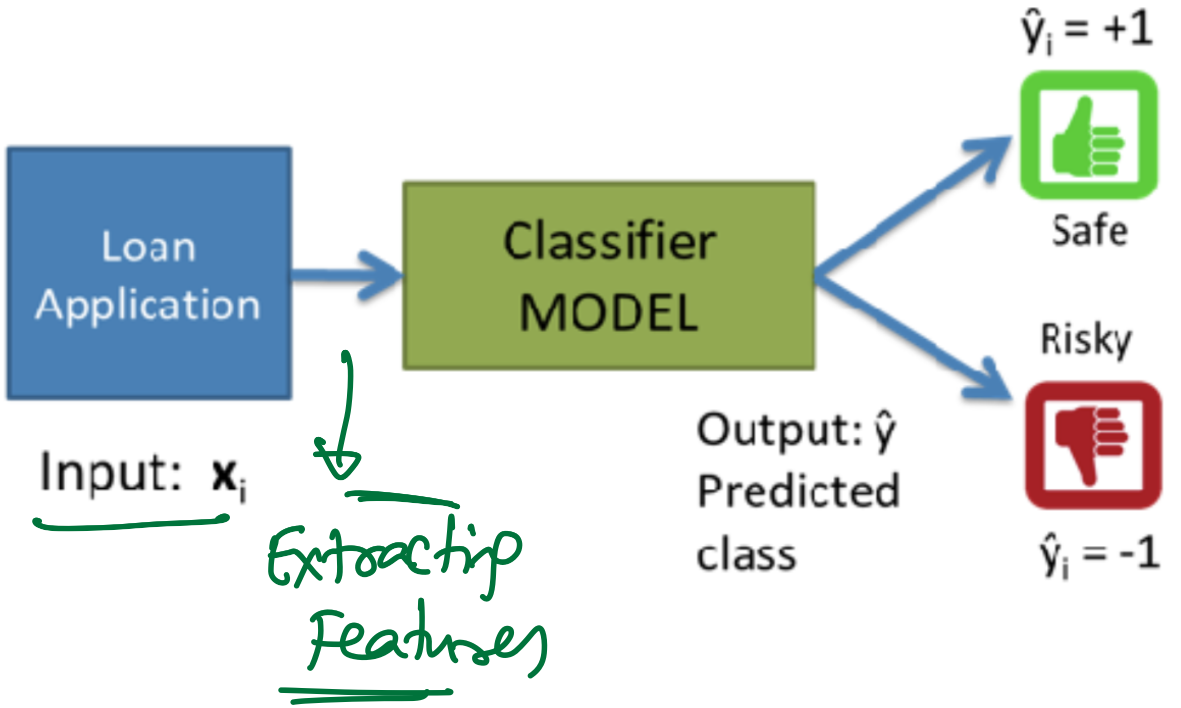
Age, reason for the loan, marital status,…

Example: Home loan for a married couple



Intelligent Loan Review System

Loan Classifier



Sample Data

Data (N observations, 3 features)

Target

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	5 yrs	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	safe
fair	5 yrs	low	safe
poor	3 yrs	high	risky
poor	5 yrs	low	safe
fair	3 yrs	high	safe

N = 9

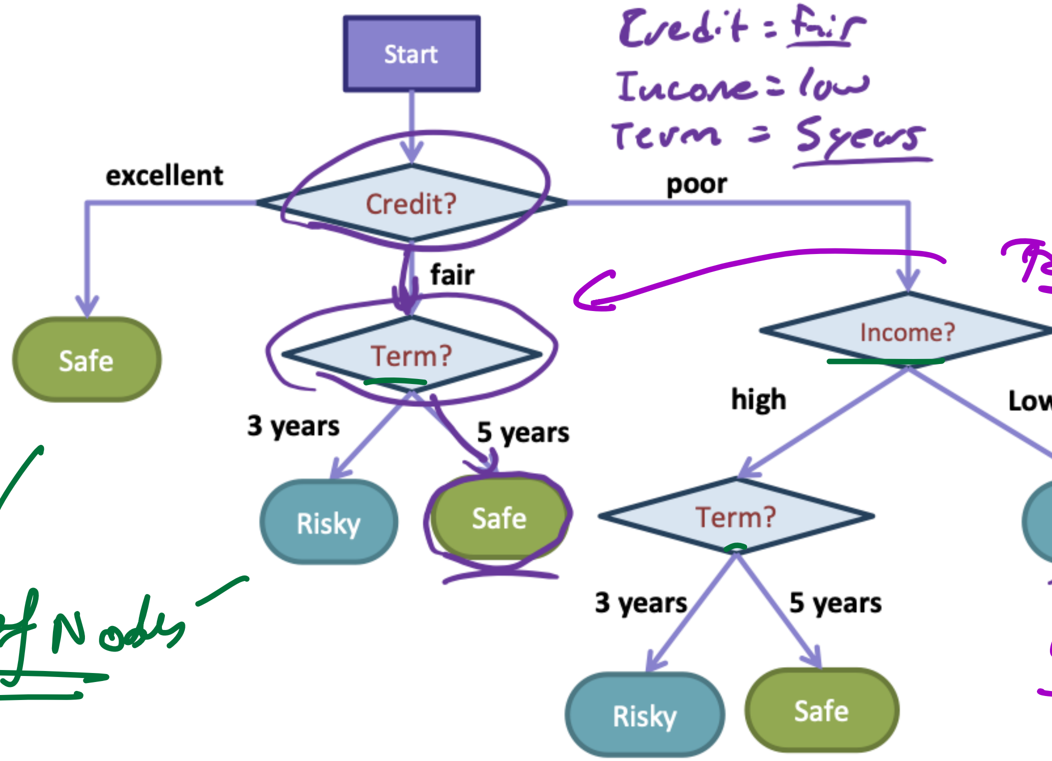
3 features

Target → Safe
Target → Risky

Decision Trees

Root Node

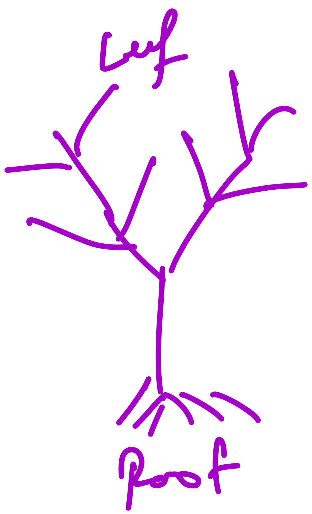
Credit = fair
Income = low
Term = 5years



Tree Node

Leaf Nodes

Leaf



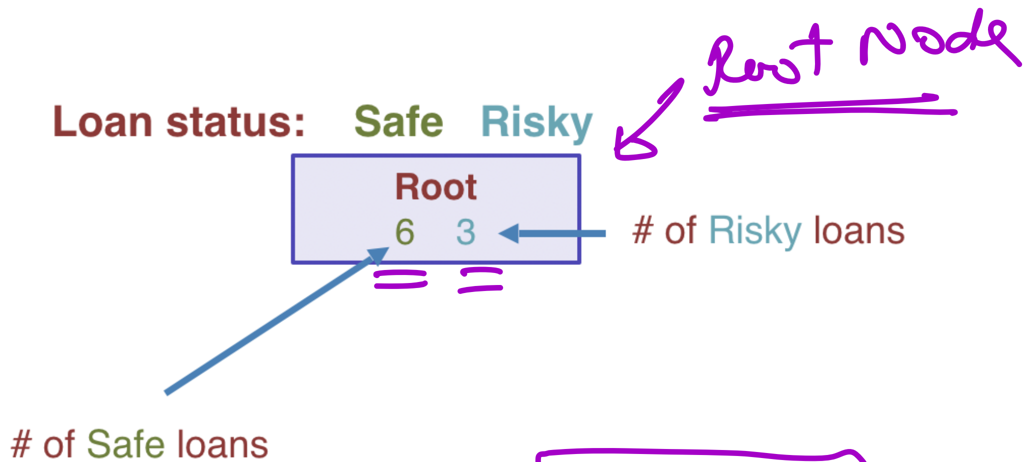
- **Branch/Internal node:** splits into possible values of a feature
- **Leaf node:** final decision (the class value)

Growing Trees

Questions

- Which features are "good"? | Growing criteria
- When to stop growing a tree? | Stopping criteria

Visual Notation

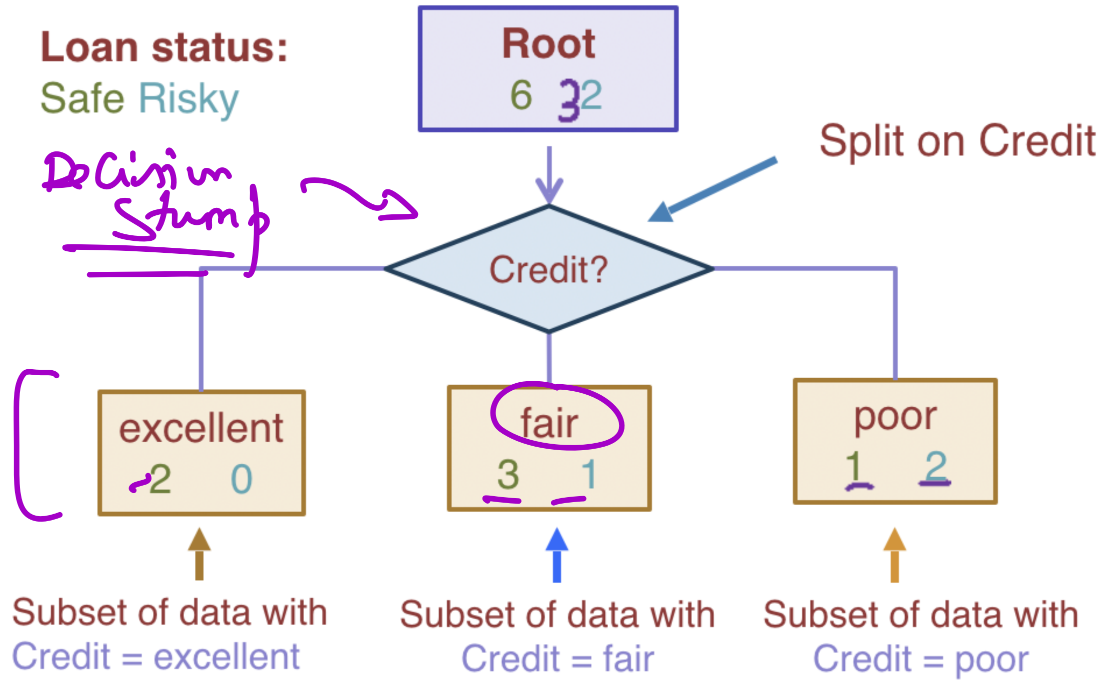


N = 9 examples

Decision stump 1

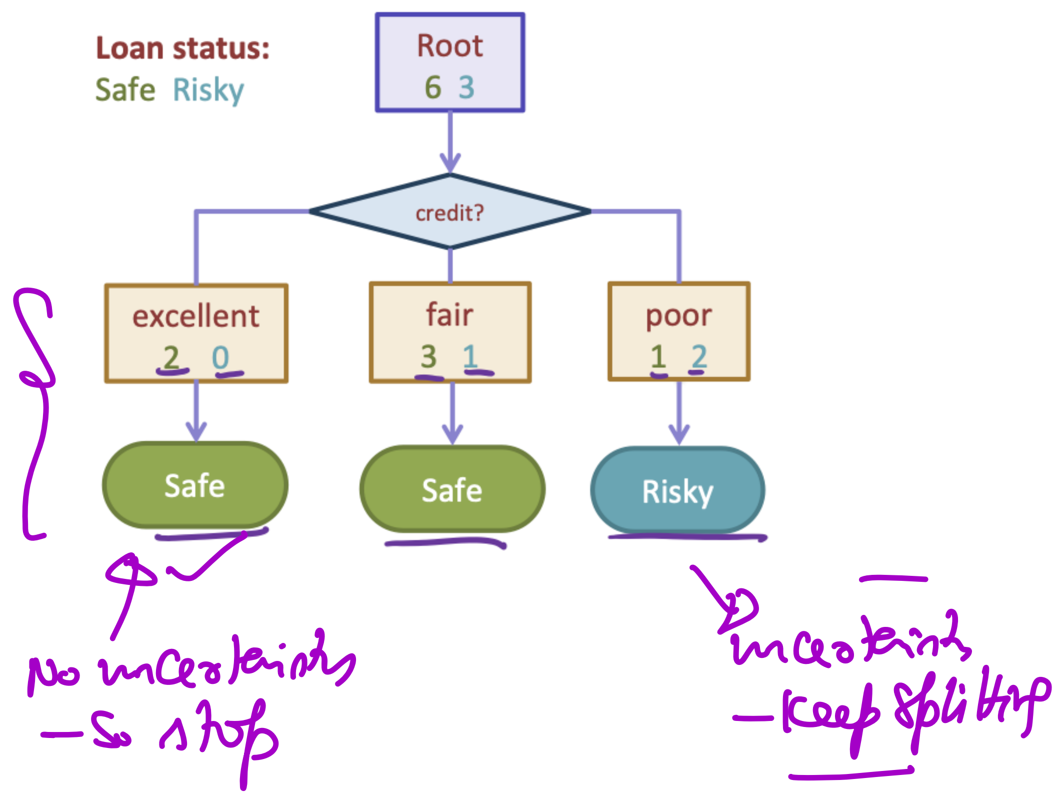
Data (N observations, 3 features)

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	5 yrs	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	safe
fair	5 yrs	low	safe
poor	3 yrs	high	risky
poor	5 yrs	low	safe
fair	3 yrs	high	safe



Making predictions

For each leaf node, set ŷ = majority value

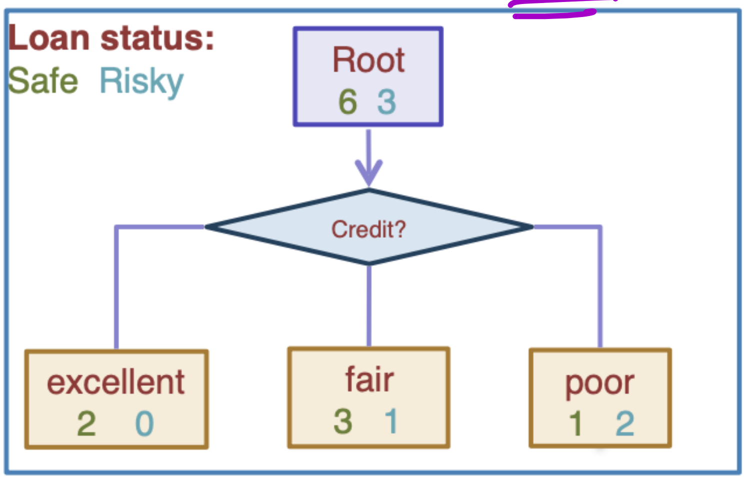


Split selection

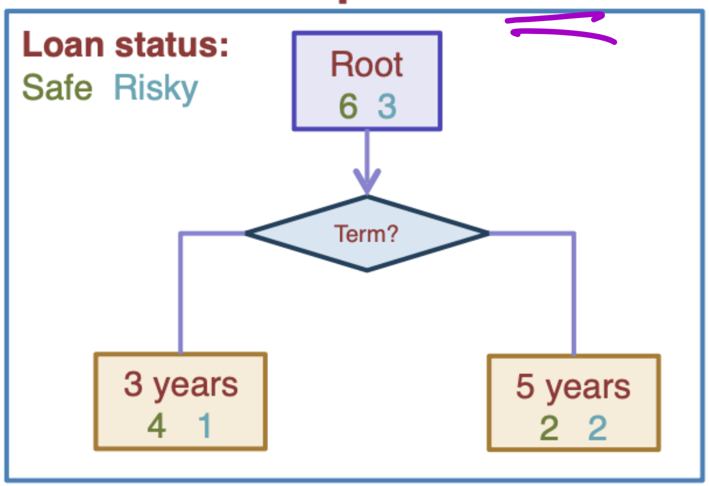
How do we select the best feature?

Select the split with lowest classification error

Choice 1: Split on Credit

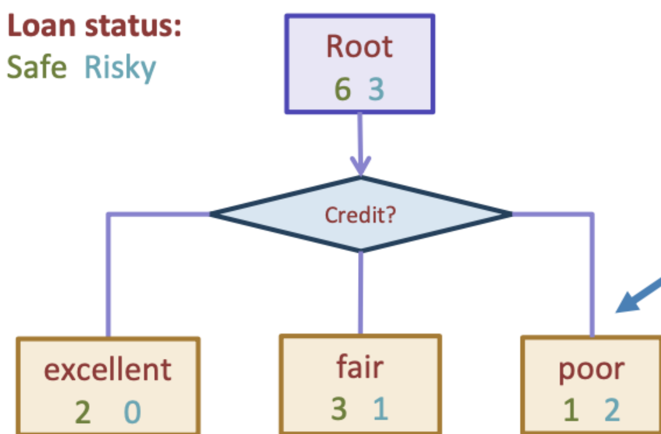


Choice 2: Split on Term



Split Effectiveness

How do we measure effectiveness of a split?



SAFE SAFE RISKY

Mistakes: 0 1 1

Idea: Calculate classification error of this decision stump

$$\text{Error} = \frac{\#\text{ mistakes}}{\#\text{ data points}}$$

Total # mistakes = 2

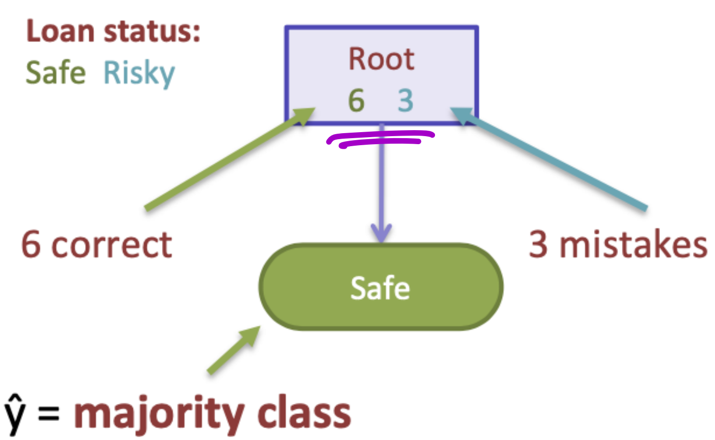
Misclassification Error = 2/9

Calculate Classification Error

Calculating classification error

Step 1: ŷ = class of majority of data in node

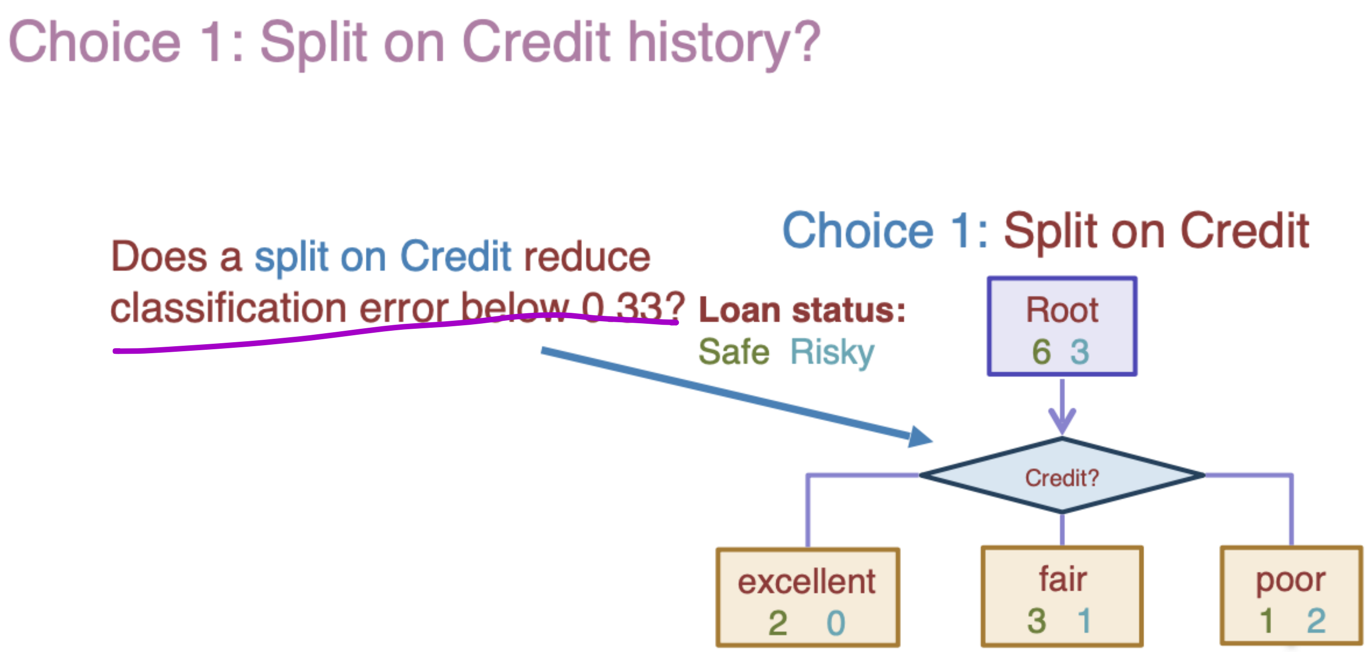
Step 2: Calculate classification error of predicting ŷ for this data



$$\text{Error} = \frac{3}{9} = 0.33$$

Tree	Classification error
(root)	0.33

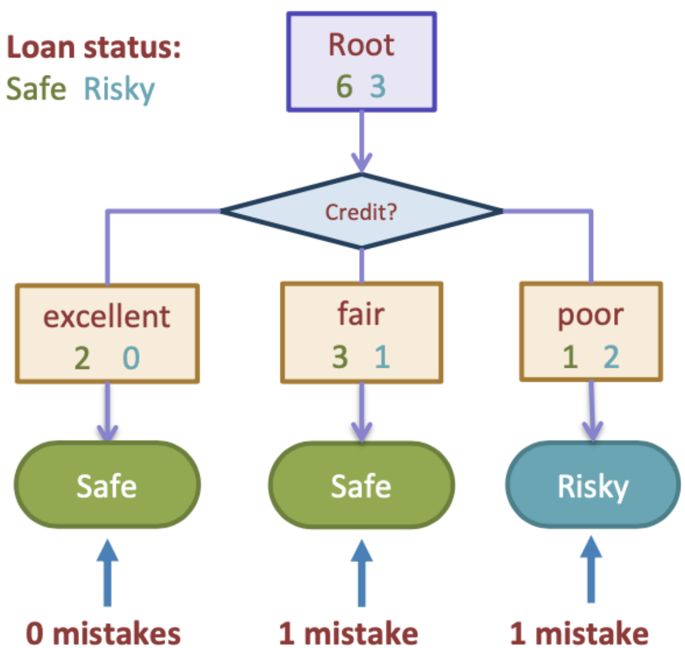
Split on Credit



Split on Credit

Split on Credit: Classification error

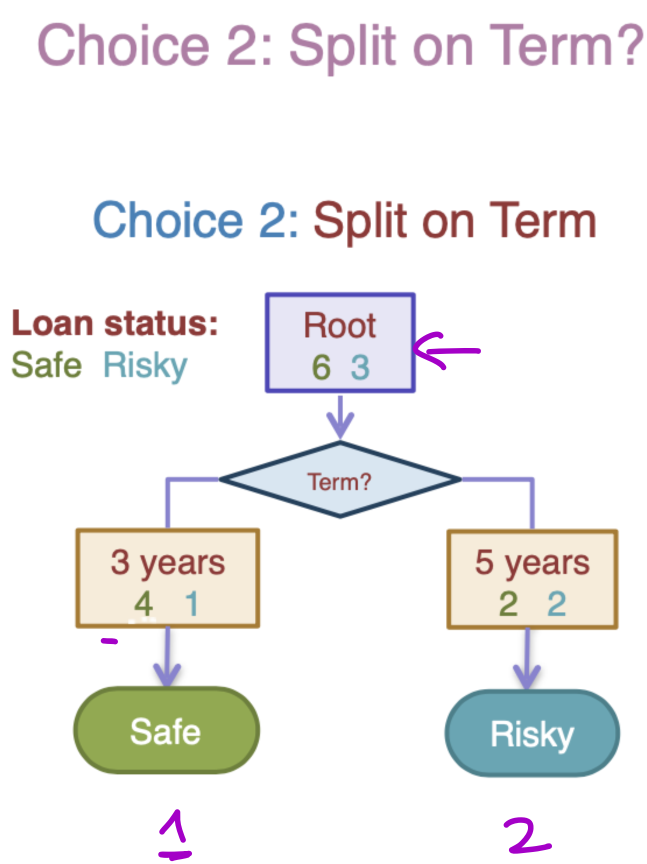
Choice 1: Split on Credit



$$\text{Error} = \frac{0+1+1}{9} = \frac{2}{9}$$
$$= 0.22$$

Tree	Classification error
(root)	0.33
Split on credit	0.22

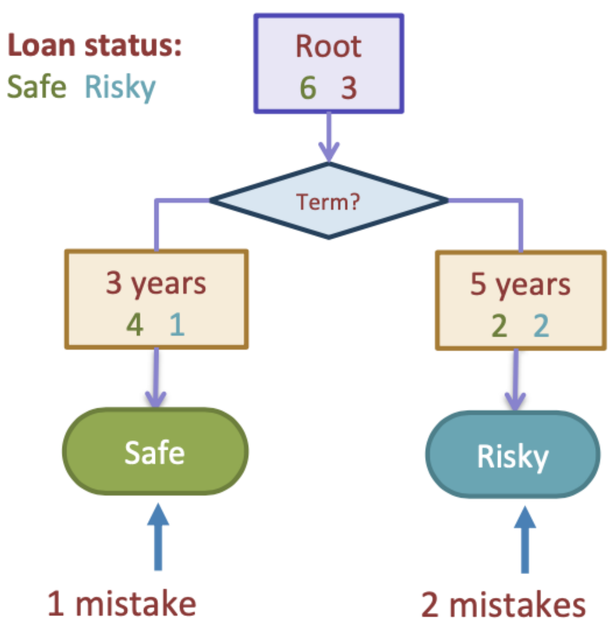
Split on Term



Split on Term

Evaluating the split on Term

Choice 2: Split on Term



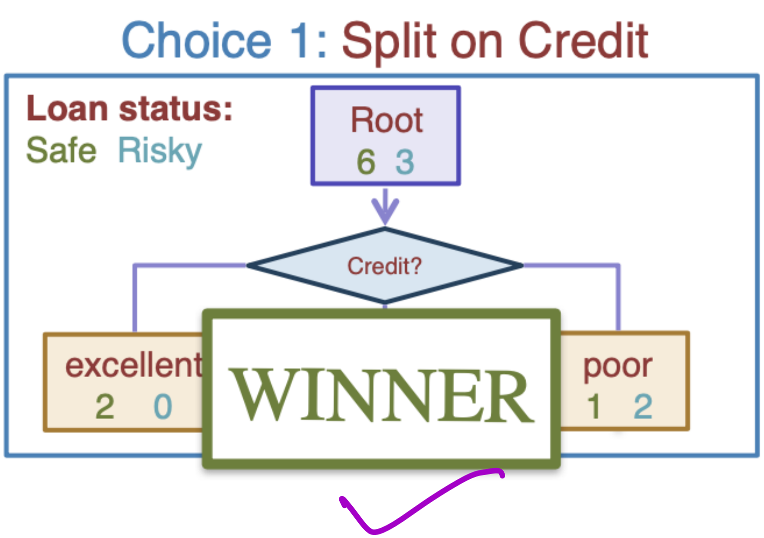
$$\text{Error} = \frac{1+2}{9} = \frac{3}{9}$$
$$= 0.33$$

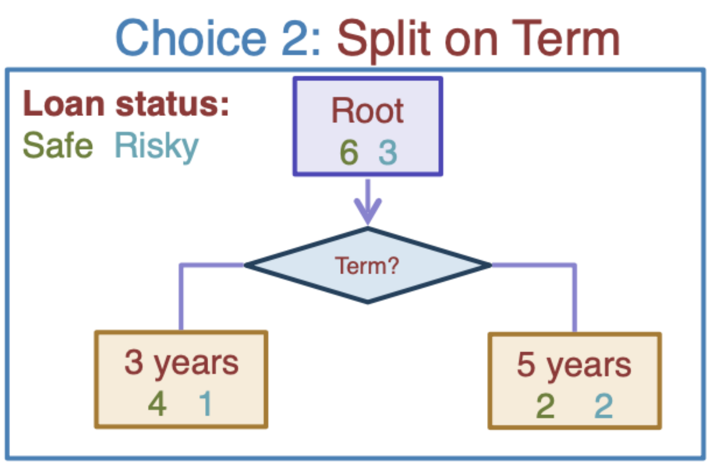
Tree	Classification error
(root)	0.33
Split on credit	0.22
Split on term	0.33

Split Winner

Choice 1 vs Choice 2: Comparing split on credit vs term

Tree	Classification error
(root)	0.33
split on credit	0.22
split on loan term	0.33





Split selection

Split selection procedure

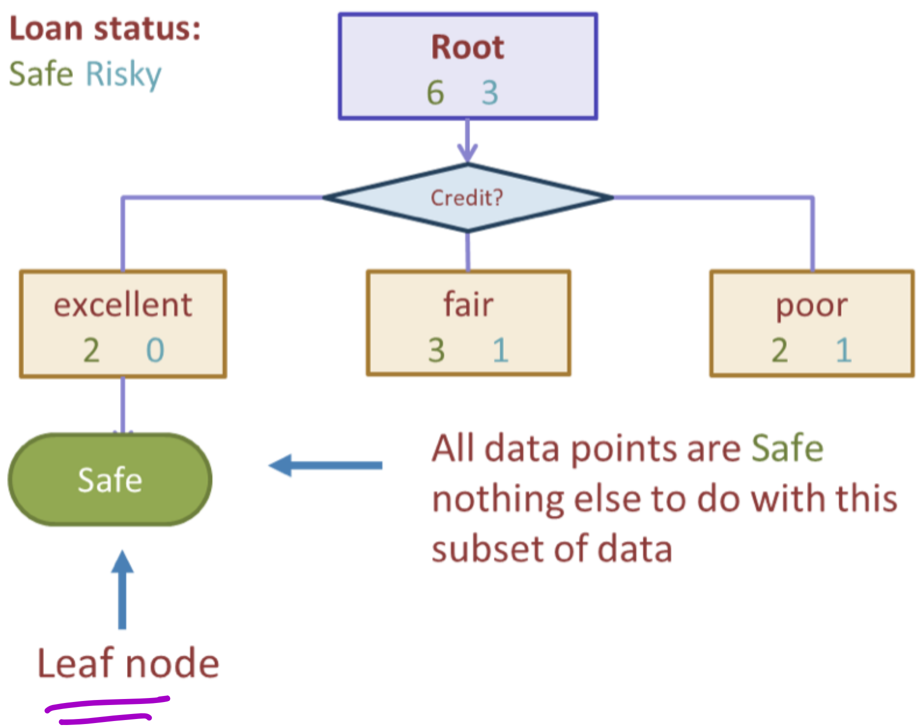
- Given a subset of data set, M at a node
- For each remaining feature $h_i(x)$, split M by feature $h_i(x)$ and compute classification error
- Pick the feature i to split with minimum classification error

Decision Tree Classification as a Greedy Procedure

DT Classifier Training procedure

If classification splits satisfy criteria (e.g. low classification error), stop,
Else, split further using split selection procedure.

Stopping



Loan status:
Safe Risky
Root
6 3
Credit?
excellent
2 0
fair
3 1
poor
2 1
Safe
All data points are Safe
nothing else to do with this
subset of data
Leaf node

Stopping criteria in practice

- **Splits with few data points** can lead to over-fitting. Example

Term

1 1

more data →

Term

30 5

Stopping criteria in practice

1. **Splits with few data points** can lead to over-fitting. Example
2. **Max tree depth** can be a stopping criteria to prevent over-fitting.

Hyper-Parameter !

Stopping criteria in practice

1. **Splits with few data points** can lead to over-fitting. Example
2. **Max tree depth** can be a stopping criteria to prevent over-fitting.
3. Although theoretically, can aim for 0 classification error - This would lead to over-fitting. Use above 2 to stop earlier.

Stopping criteria in practice

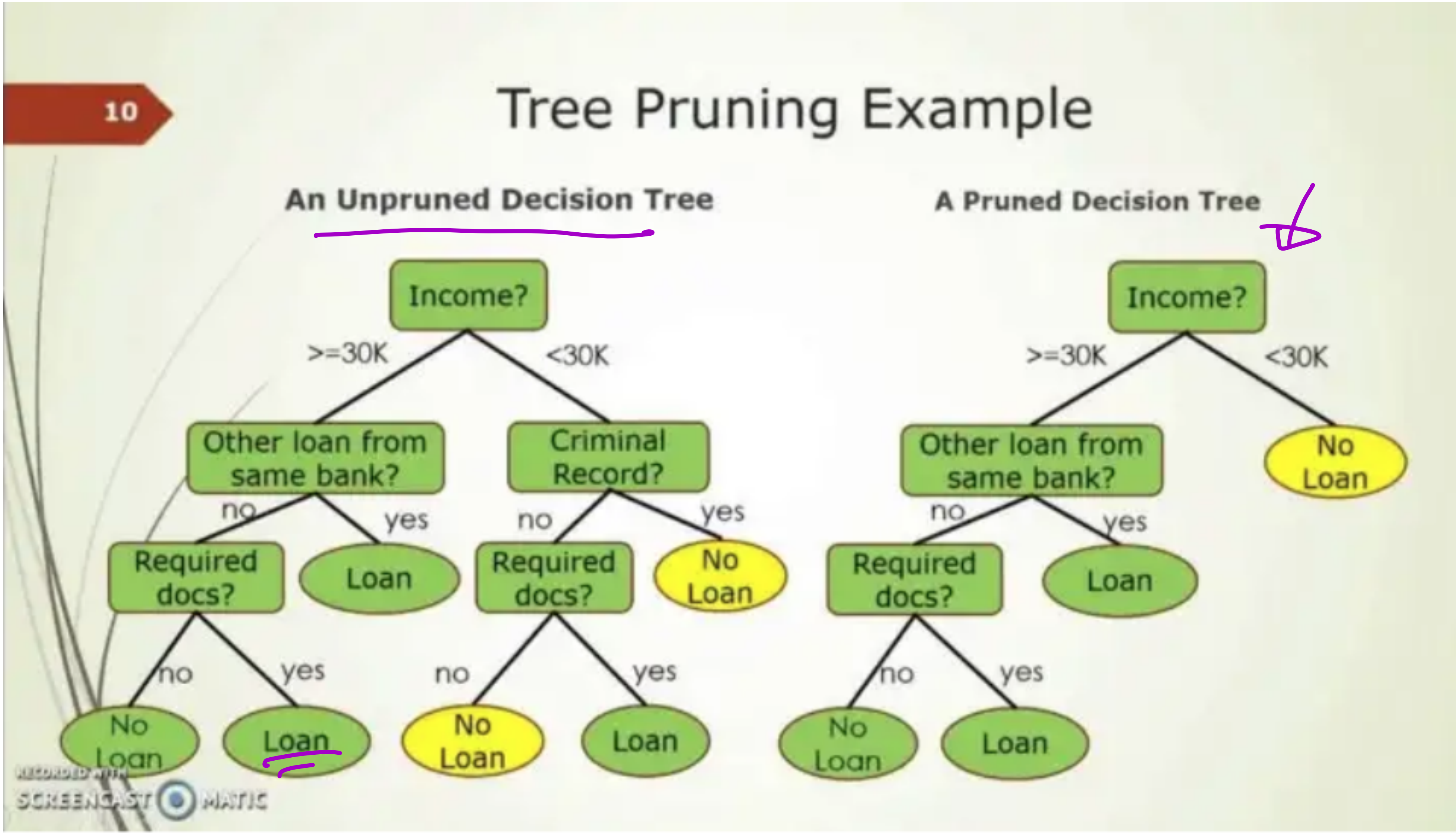
1. **Splits with few data points** can lead to over-fitting. Example
2. **Max tree depth** can be a stopping criteria to prevent over-fitting.
3. Although theoretically, can aim for 0 classification error - This would lead to over-fitting. Use above 2 to stop earlier.
4. No standard 'regularization' for DTs like for Logistic Regression. Why?

Stopping criteria in practice

1. **Splits with few data points** can lead to over-fitting. Example
2. **Max tree depth** can be a stopping criteria to prevent over-fitting.
3. Although theoretically, can aim for 0 classification error - This would lead to over-fitting. Use above 2 to stop earlier.
4. No standard 'regularization' for DTs like for Logistic Regression. Why?
5. **Pruning** - Can be done to prune branches that lead to over-fitting

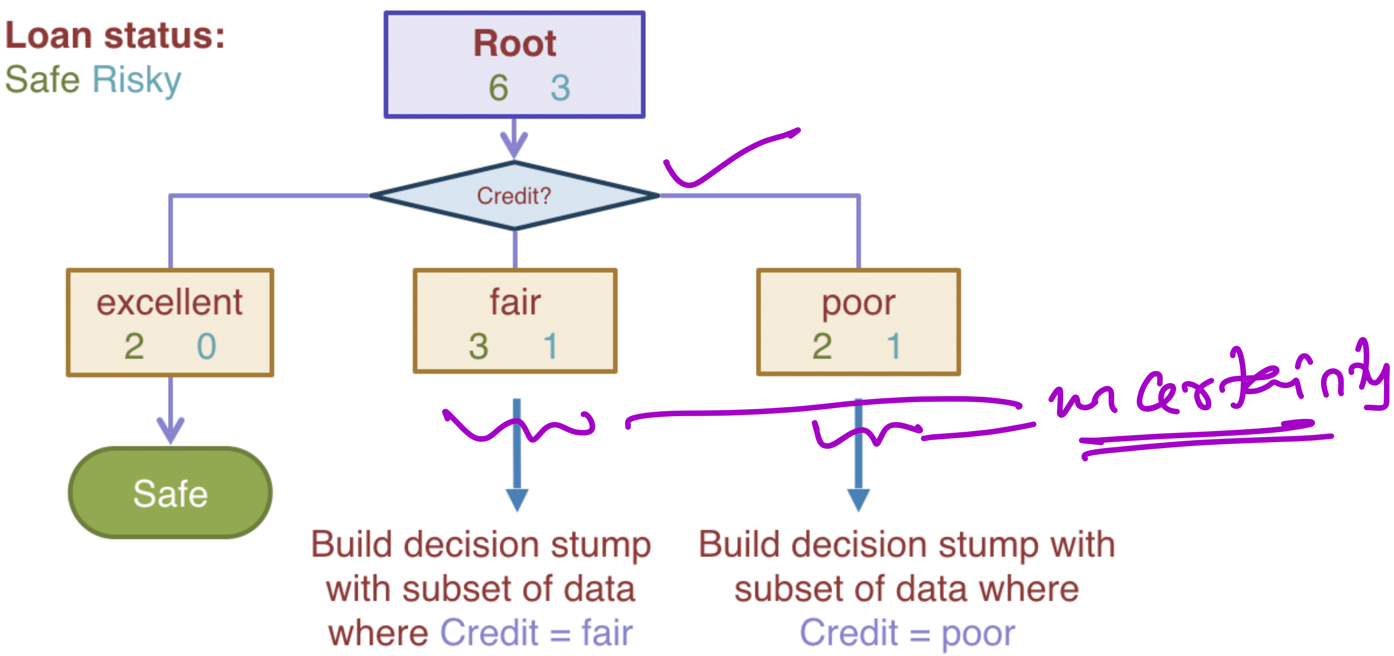
(Prune so validation error goes down!)

Decision Trees Pruning

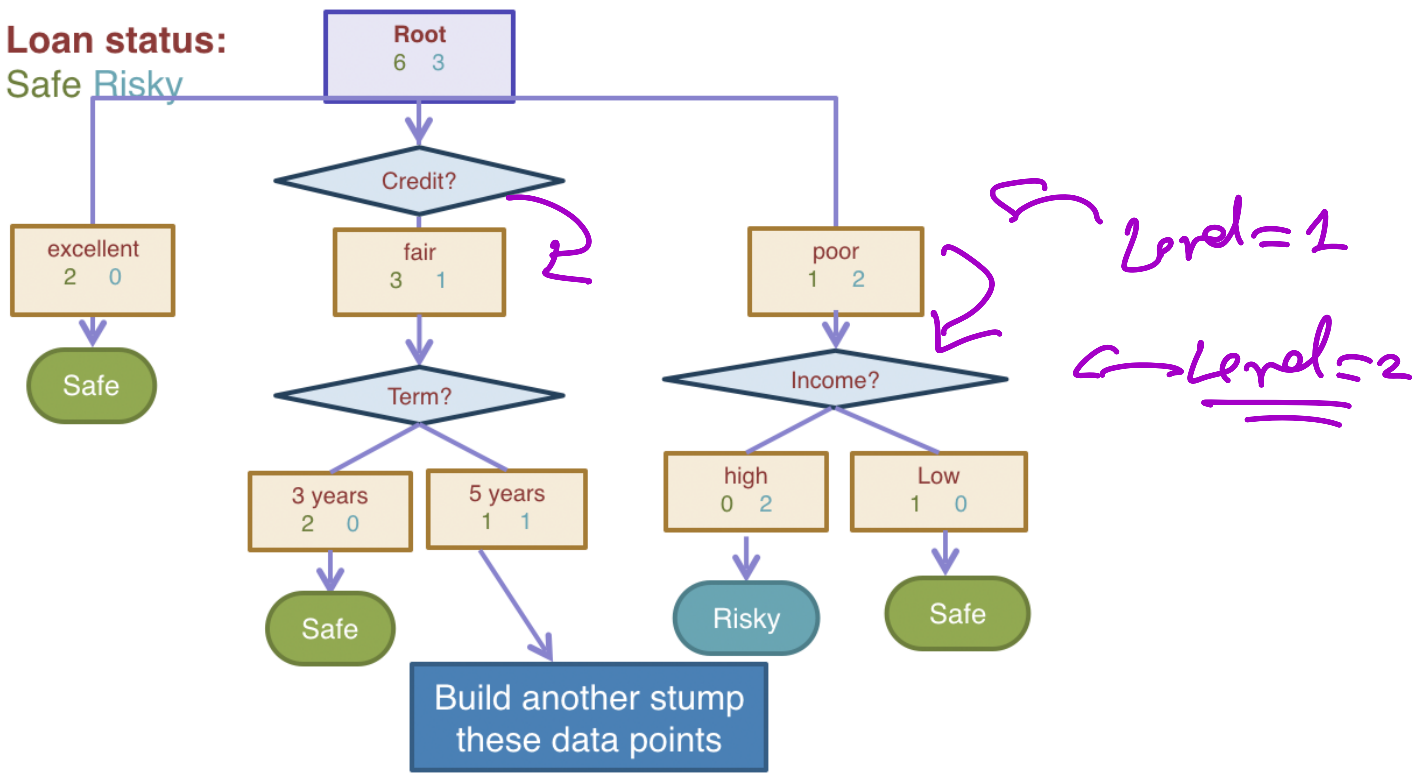


Tree Pruning Example Reference

Recursive Splits



Second level DT

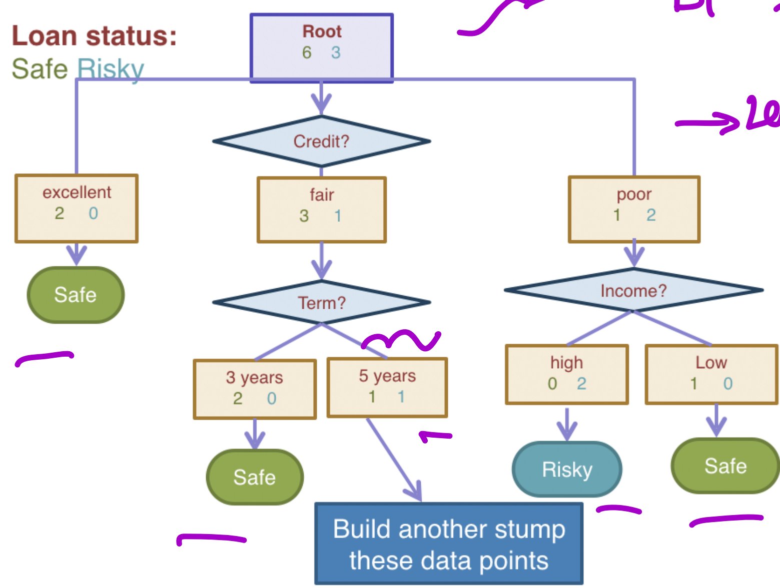


ICE #2

Classification error

Level=0
DT → 0.33

a) 0.33
b) 0.11
c) 0.22
d) 0



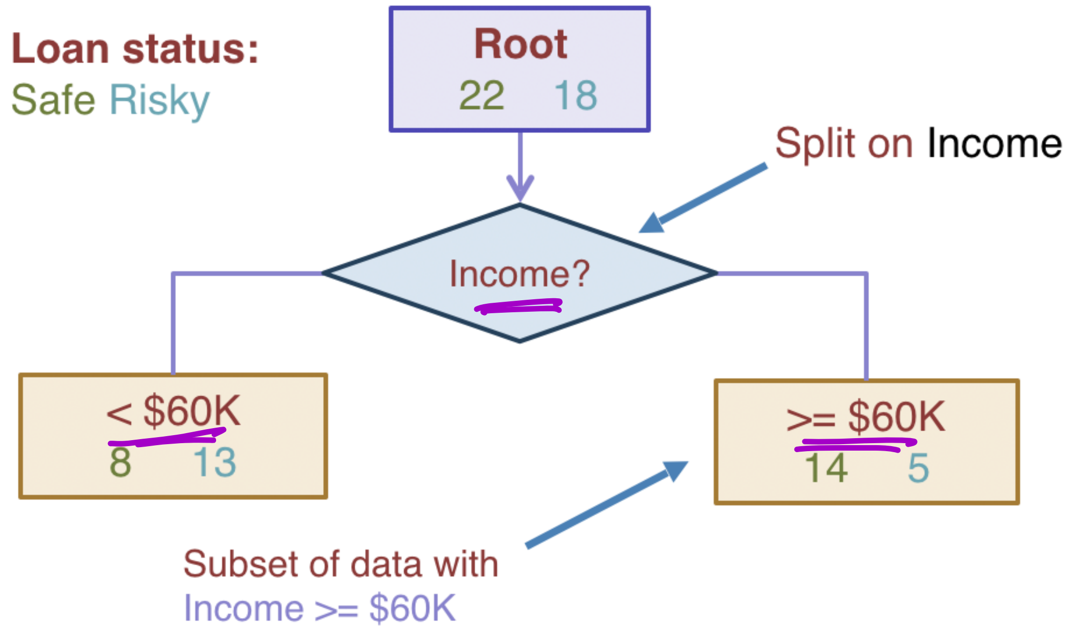
→ Level=1
DT → 0.22

→ Level=2
DT → 0.11

The classification error for the DT above is:

a) 0.33
b) 0.11
c) 0.22
d) 0

Threshold splits for real valued features



Real-valued Features:- need to pick the threshold 't'

Ex:- Income < t   Income > t

30k, 60k, 100k, 200k

Choosing Split Threshold for Numeric Features

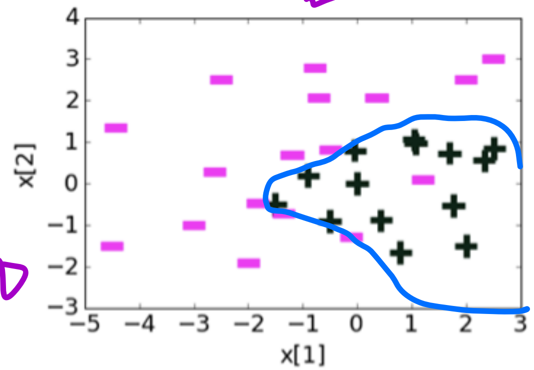
1. Grid search?
2. Numeric vs Categorical Features: Can recurse more than once on a numeric feature. Can't do the same for categorical feature. Why?

Income > 60k → Yes → Income < 150k → No

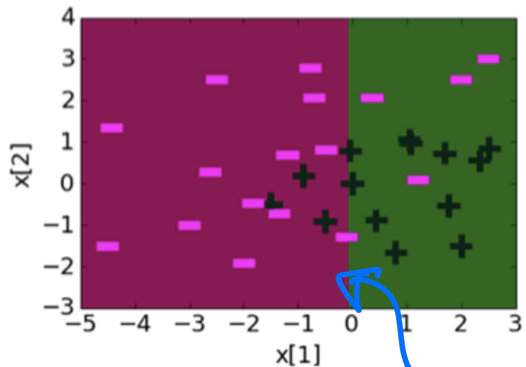
Decision Boundary level 1 || Numeric Features

31 data pts

Before split



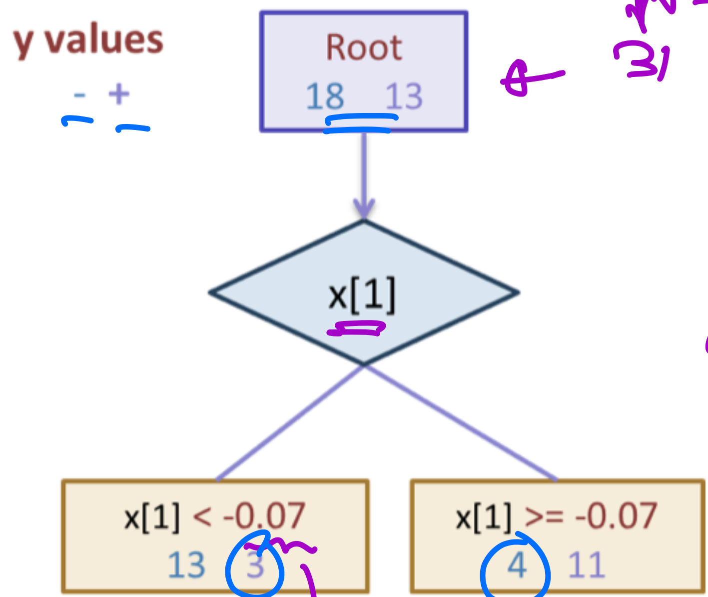
After split



"vertical separation boundary"

classification error = 13/31

N = 31

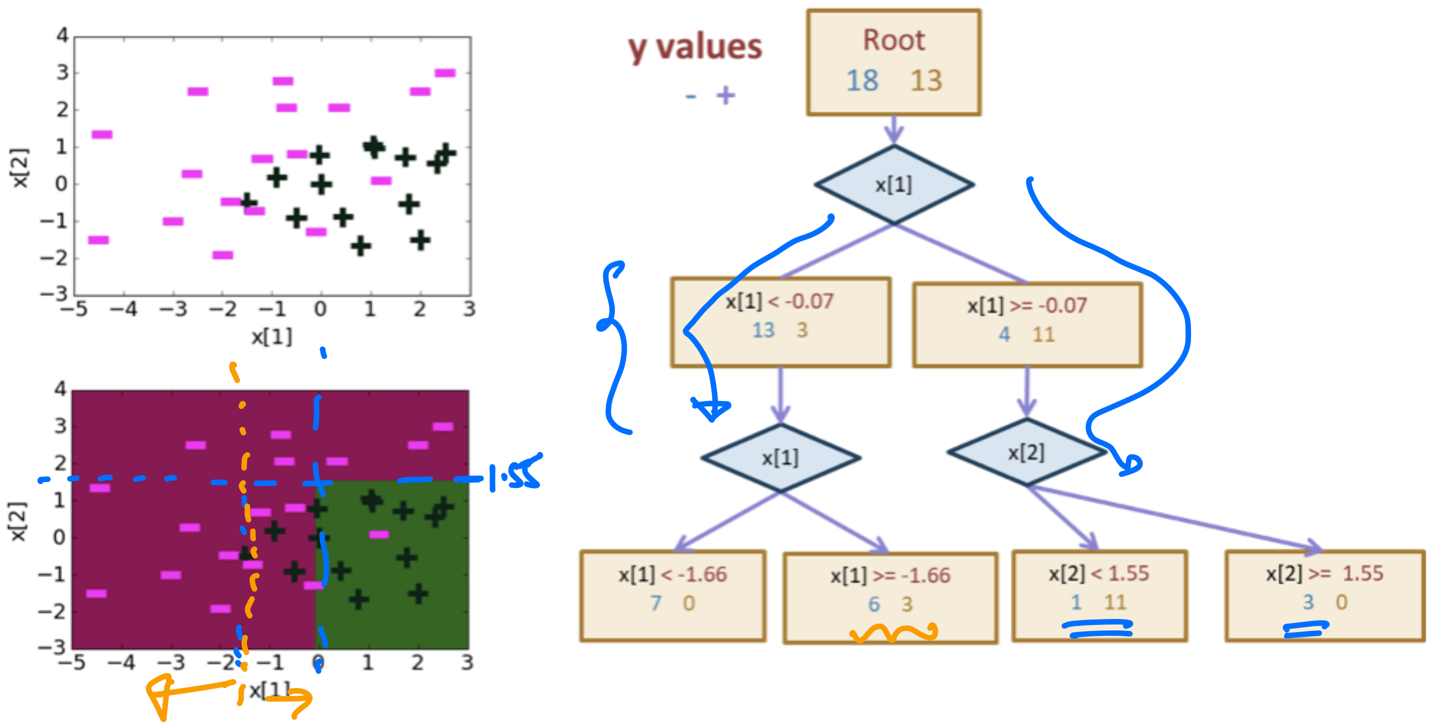


t = -0.07

Both x[1] & x[2] are 'Numeric' Features

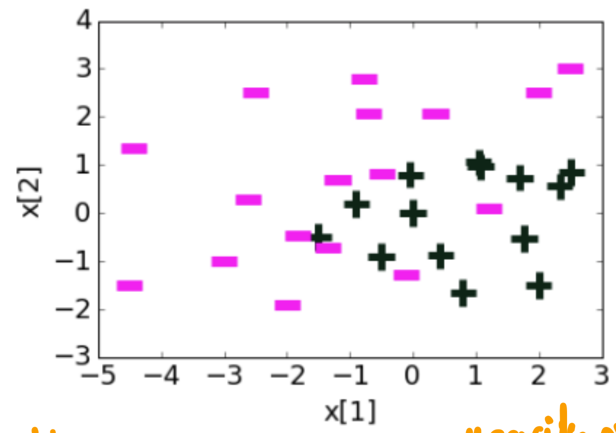
classification error = 7/31

Decision Boundary level 2 || Numeric Features



Decision Boundary level 3 || Numeric Features

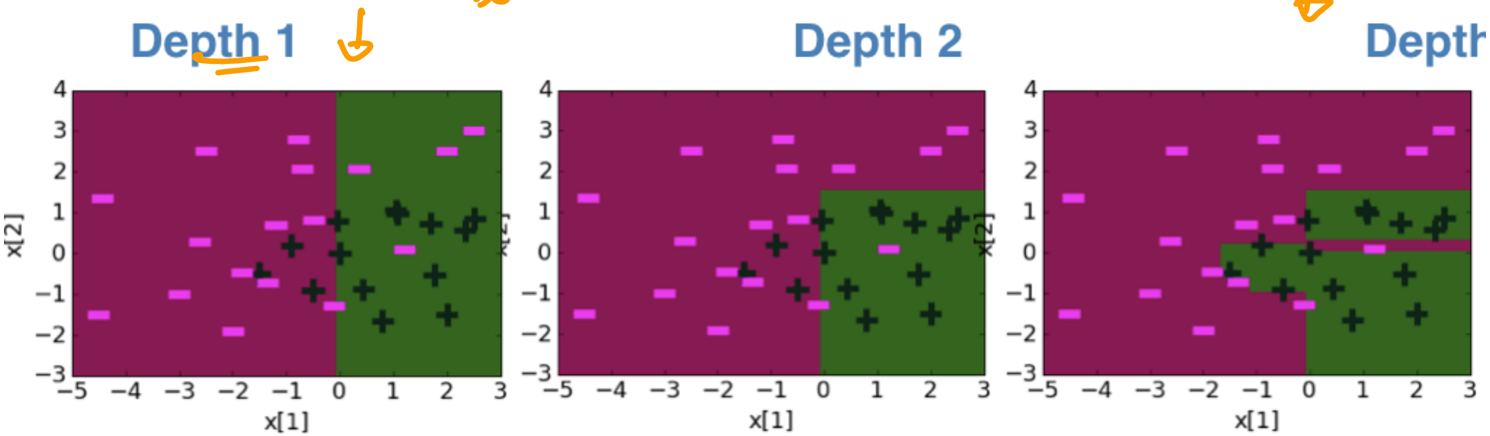
- Decision boundaries can be complex!



Linear Decision boundary

non-linear

Heavily non-linear



Decision Trees Summary

Summary

- Intuitive way to classify by making decsisions by walking down the tree
- Can learn complex **non-linear decision boundaries** (unlike logistic regression)
- Prone to **overfit** as tree depth increases (unlike logistic regression)
- Splitting at nodes with few data points can lead to overfitting
- Over-fitting can be avoided by early stopping (depth or error)
- Improve Decision Trees - Random Forests - Next Lecture!

Decision Trees vs Logistic Regression

1. Both are **interpretable** in different ways

Decision Trees vs Logistic Regression

1. Both are **interpretable** in different ways
2. Decision trees **mimick how humans make decisions** and are useful in certain contexts - Like medical diagnosis or other places where number of features is not too large

Decision Trees vs Logistic Regression

1. Both are **interpretable** in different ways
2. Decision trees **mimick how humans make decisions** and are useful in certain contexts - Like medical diagnosis or other places where number of features is not too large
3. Decision Trees can easily learn **non-linear decision boundaries** while Logistic Regression learns linear decision boundary

Decision Trees vs Logistic Regression

1. Both are **interpretable** in different ways
2. Decision trees **mimick how humans make decisions** and are useful in certain contexts - Like medical diagnosis or other places where number of features is not too large
3. Decision Trees can easily learn **non-linear decision boundaries** while Logistic Regression learns linear decision boundary
4. Decision Tree has a higher model complexity as compared to Logistic Regression

Decision Trees vs Logistic Regression

1. Both are **interpretable** in different ways
2. Decision trees **mimick how humans make decisions** and are useful in certain contexts - Like medical diagnosis or other places where number of features is not too large
3. Decision Trees can easily learn **non-linear decision boundaries** while Logistic Regression learns linear decision boundary
4. Decision Tree has a higher model complexity as compared to Logistic Regression
5. Logistic Regression is less prone to **over-fitting** than Decision Trees with large number of features

Pitfalls of Decision Trees

1. **Overfitting**

Pitfalls of Decision Trees

1. **Overfitting**
2. **Feature Engineering**

Pitfalls of Decision Trees

1. **Overfitting**
2. **Feature Engineering**
3. **Not suitable for Regression**

Overcoming pitfalls of Decision Trees - Random Forests

Random Forests Introduction

A **Random Forest** is a collection of T Decision Trees. Each decision tree casts a "vote" for a prediction and the ensemble predicts the majority vote of all of its trees.