

Lecture 7: Classification Models and Cross Validation

INFO 1989 (Taylor's Version): Introduction to Machine Learning

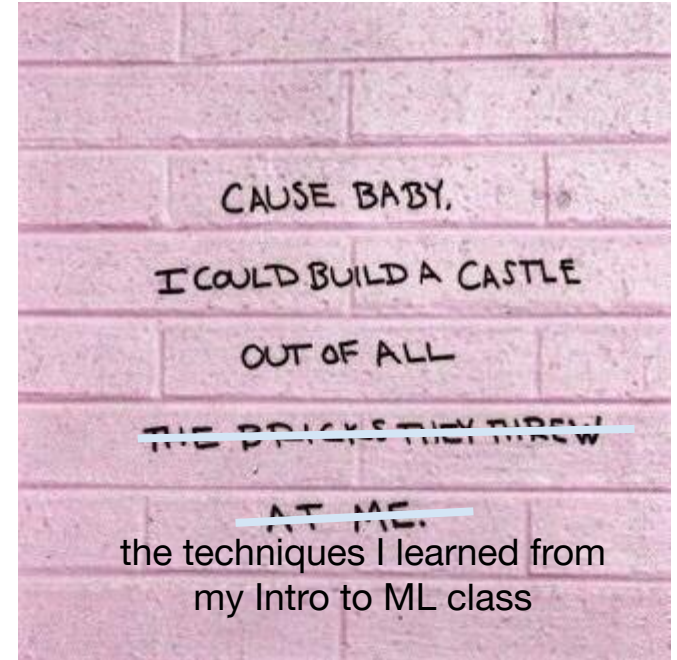


CDS Education

We explore, learn, and educate big minds.

Agenda

1. Decision Trees
2. Logistic Regression and Its Applications
3. Cross Validation



Review: Classifiers



What are Classifiers?

- Classifiers predict the class/category of a set of data points. This class/category is based off of the target variable we are looking at.
- Examples:
 - “What species is this?”
 - “What major is a student in based on their classes?”

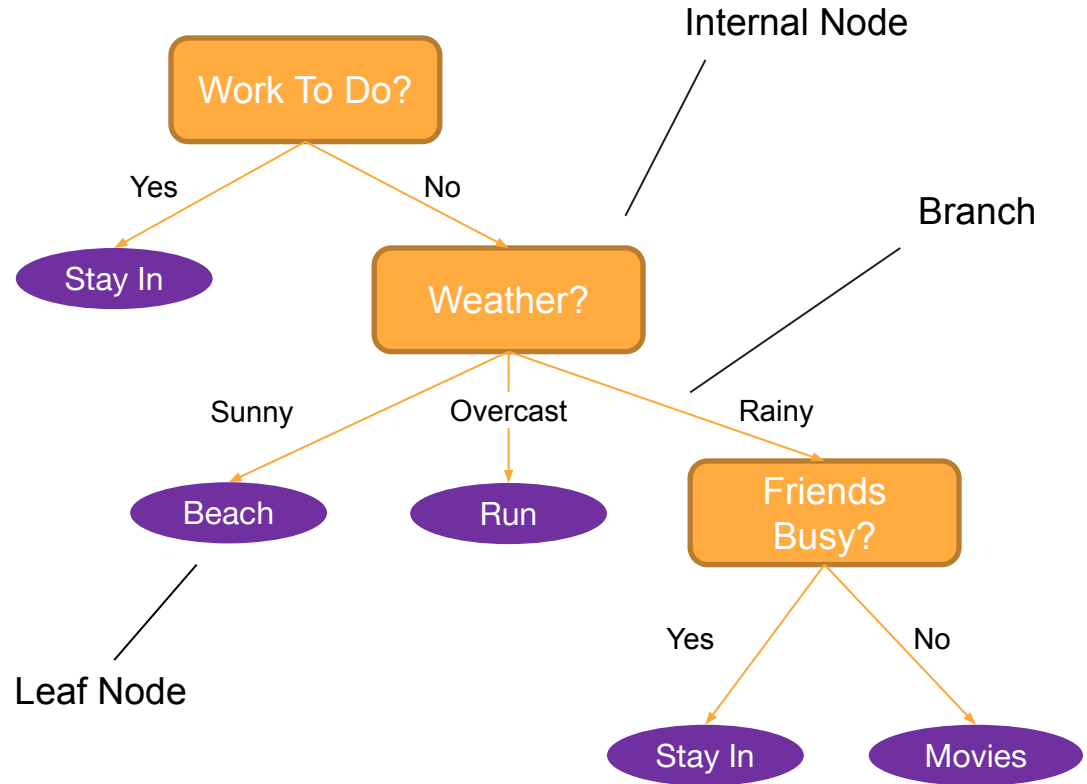


Decision Trees



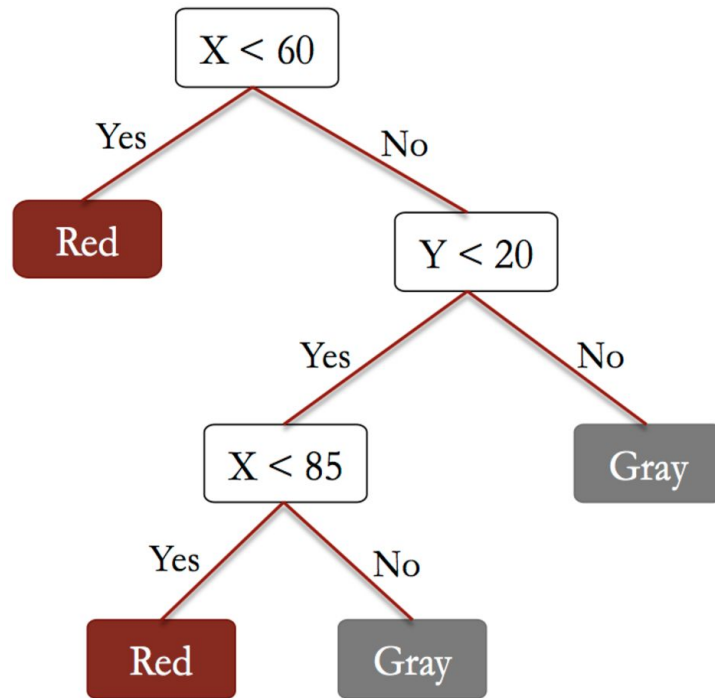
How Should I Spend My Weekends

A **decision tree** is a supervised machine learning model used to predict a target by learning decision rules from features. As the name suggests, we can think of this model as **breaking down** our data by **making a decision** based on **asking a series of questions**.



CART (Classification and Regression Trees)

- Used for Classification and Regression
- At each node, split on variables
- Each split minimizes error function
- Very interpretable
- Models a non-linear relationship!



Pros and Cons of Using Decision Trees

Pros	Cons
Easy to interpret	Overfitting 😞
Requires little data preparation (robust to missing data)	Requires parameter tuning (max depth)
Can use a lot of features	
Can capture non-linear relationships	



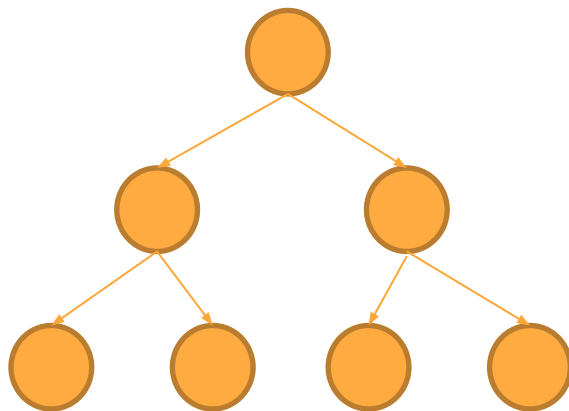
How to Reduce Overfitting

1. Limit the max depth of the tree

Depth = 0

Depth = 1

Depth = 2



Model
Complexity

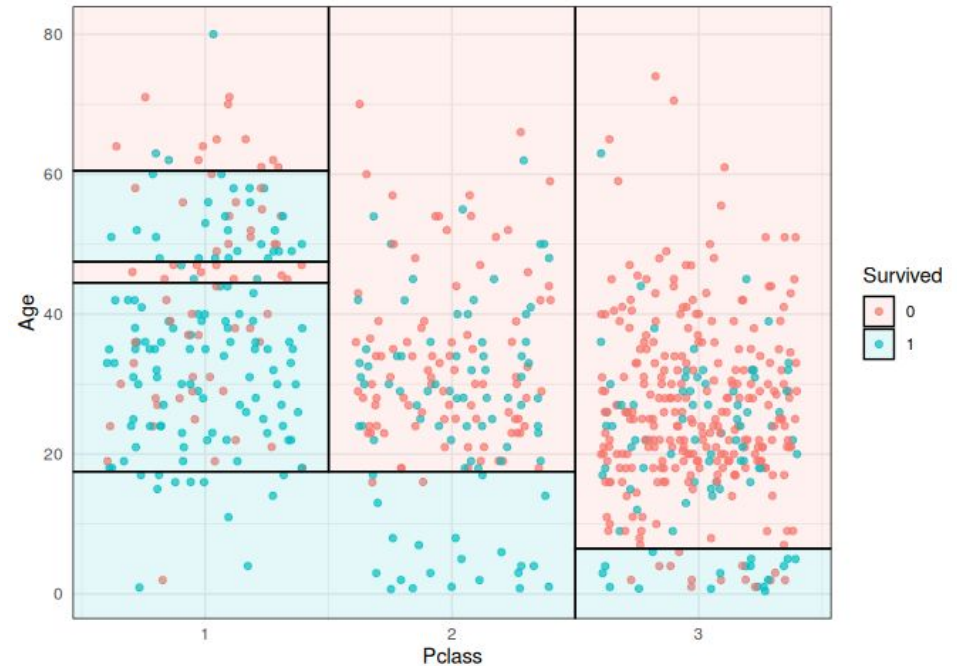


When training a decision tree, we have to specify the maximum depth a constructed tree can have



How to Reduce Overfitting

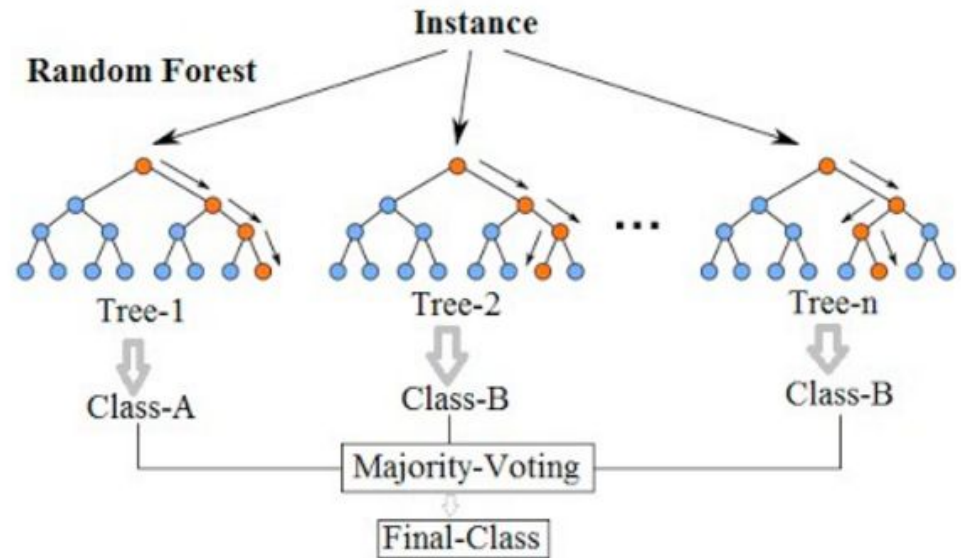
- There are no “curves” for each decision tree boundary line
- Limiting the depth of the tree limits the number of lines you are splitting on



How to Reduce Overfitting

2) Train multiple decision trees and determine final output based on output of each decision tree

This is called a
Random Forest Classifier



Demo



Logistic Regression



Logistic Regression

- Used for Binary Classification:

$$Y = \begin{cases} 1 \\ 0 \end{cases}$$

- Fits a linear relationship between the variables
- Transforms the linear relationship of probability that the outcome is 1 by using the **sigmoid function**

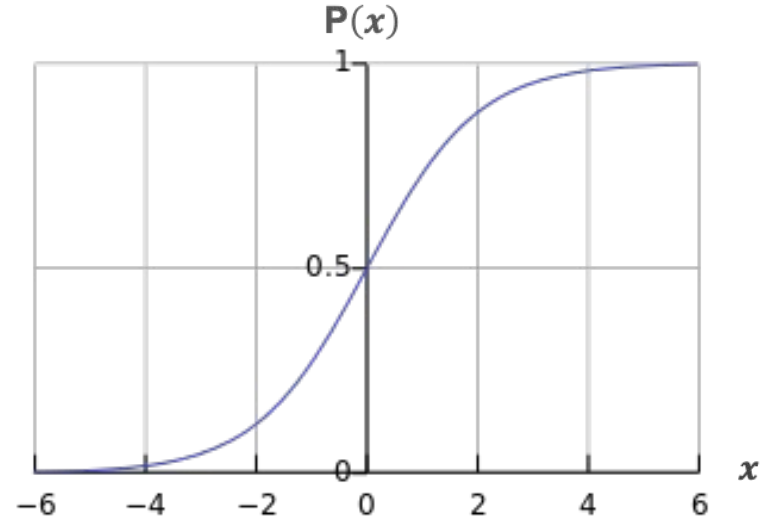
Formula:

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}} \longrightarrow \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$



Logistic Function

$$\hat{P}(x) = \frac{1}{1 + e^{-x}}$$



The Logistic Function “**squeezes**” numbers to be between 0 and 1



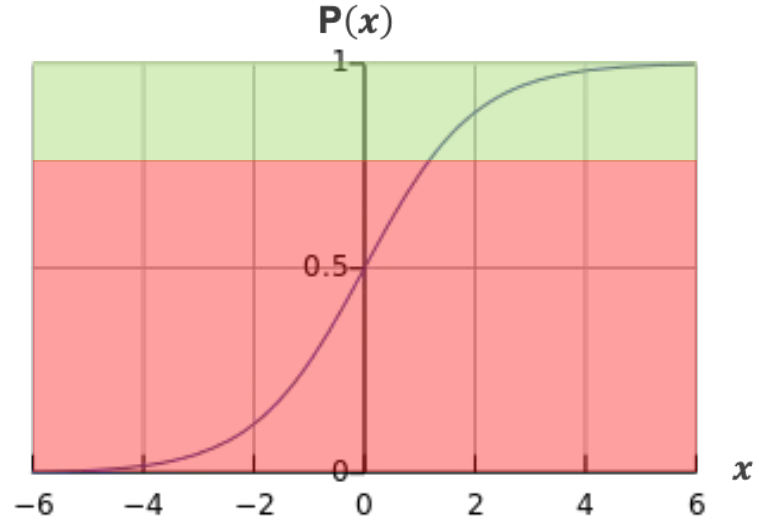
Allows us to interpret our prediction as a “**probability**” that something is true



Threshold

At what point do we differentiate between our classifications?

- $f(x)$ **below** threshold: predict 0
- $f(x)$ **above** threshold: predict 1



Pros and Cons of Using Logistic Regression

Pros	Cons
Easy to interpret (probability)	Only Capable of Binary Classification
Computationally efficient to compute	
Does not require parameter tuning	

Logistic Regression is a simple model, therefore, oftentimes it is used as a good “baseline” to compare more complex models to



Demo



Cross Validation



What is Cross Validation?

- Cross validation is a resampling technique that uses different parts of a dataset to assess the performance of a model
- We will look at two methods:
 - K-Fold
 - Leave-One-Out



K-fold Cross Validation



Often used in practice with $k=5$ or $k=10$.

Create equally sized k partitions, or **folds**, of training data

For each fold:

- Treat the $k-1$ other folds as training data.
- Test on the chosen fold.

The average of these errors is the validation error



***K*-fold Cross Validation**

Dataset

**Suppose $K = 5$,
5-Fold CV**



***K*-fold Cross Validation**

Fold 1

Fold 2

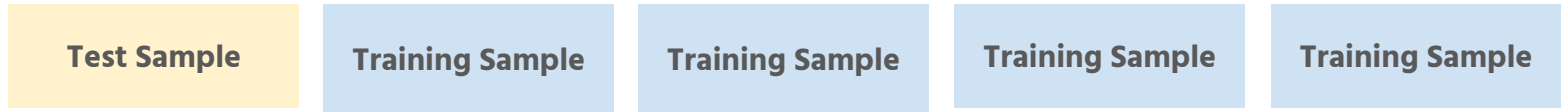
Fold 3

Fold 4

Fold 5



K-fold Cross Validation



Calculate $MSE = mse_1$



***K*-fold Cross Validation**

Training Sample

Test Sample

Training Sample

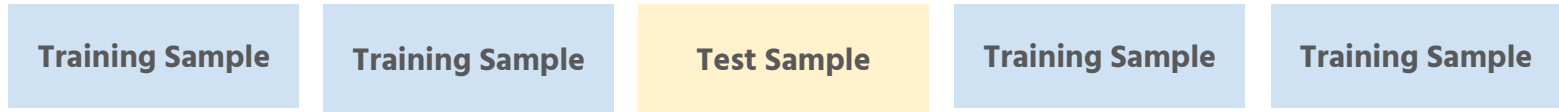
Training Sample

Training Sample

Calculate $MSE = mse_2$



K-fold Cross Validation



Calculate $MSE = mse_3$



***K*-fold Cross Validation**

And so on



K-fold Cross Validation

Fold 1

Fold 2

Fold 3

Fold 4

Fold 5

$$\text{MSE} = \text{Avg}(\text{mse1...5})$$



***K*-fold Cross Validation**

**Matters less
how we divide
up**

**Selection bias
not present**



Leave-1-Out Cross Validation

For each sample:

- Treat all other data as training data.
- Test on that one sample

The average of these errors is the validation error

Pro: Better on small datasets

Pro: More realistic (trained on most of the data)

Con: Takes longer to run



Demo



Coming Up

- **Assignment 7:** Due on November 1st, 2023
- **Next Lecture:** Linear Classifiers and Model Validation
- **Last Day for Mid-Semester Check In:** November 1st, 2023

This lecture when you're working on A7:



CDS Education

We explore, learn, and educate big minds.