Week 06: Machine Learning - II

Data Science Bootcamp
Fall, 2021

Instructor: Sagar Patel

Halfway there...

Introduction

Introduction to Python; Getting started with Git and GitHub

SQL

Relational databases; Selecting, joining and updating tables using SQL

Python Libraries

Python packages and libraries; Working with numpy, pandas, matplotlib and seaborn

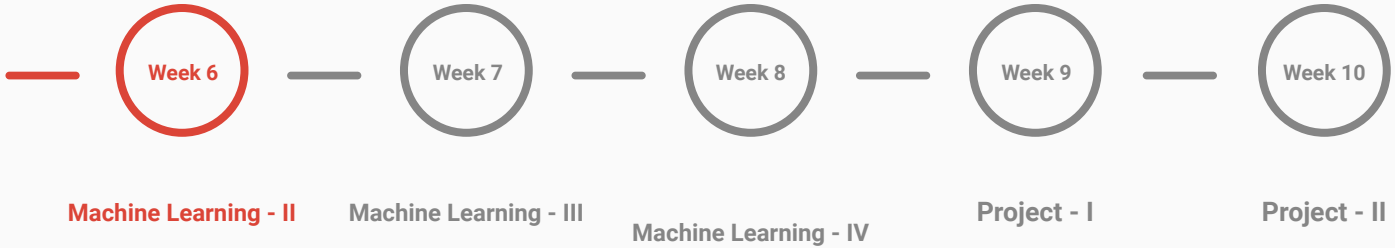
Exploratory Data Analysis

What makes data good? The “recipe” for performing EDA

Machine Learning - I

Introduction to Machine Learning; Types of Learning

Halfway there...



Agenda

- Supervised and Unsupervised Learning
  - Classification; Regression (Linear and Logistic)
  - Clustering
- Model Representation
- Confusion Matrix

A quick Recap

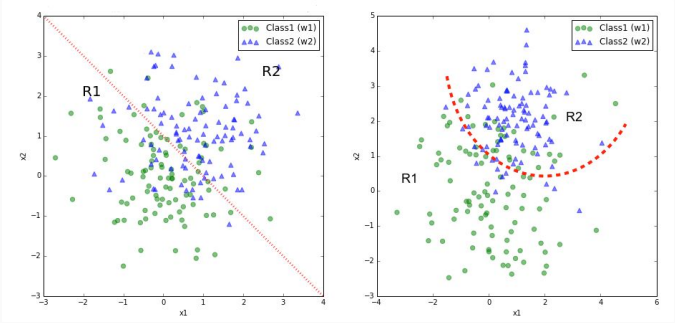
slido

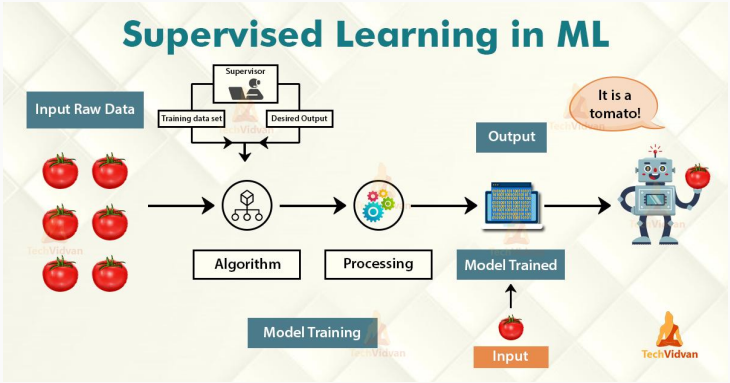
Tomato Detector

ⓘ Start presenting to display the poll results on this slide.

Types of “Learning”

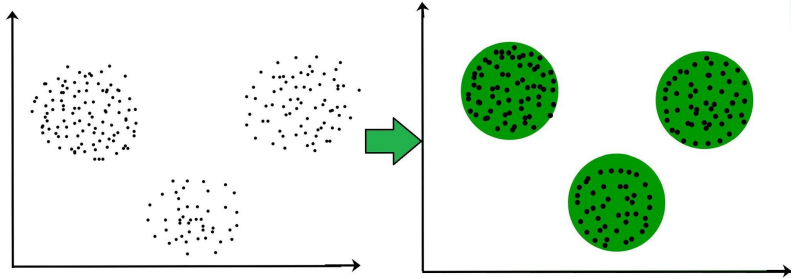
- **Supervised Learning**
  - Algorithms trained using **labeled** data
  - The model takes direct **feedback** to check if it’s predicting the correct output or not
  - Can be categorized in **Classification** and **Regression**
  - Example: Tomato Detector

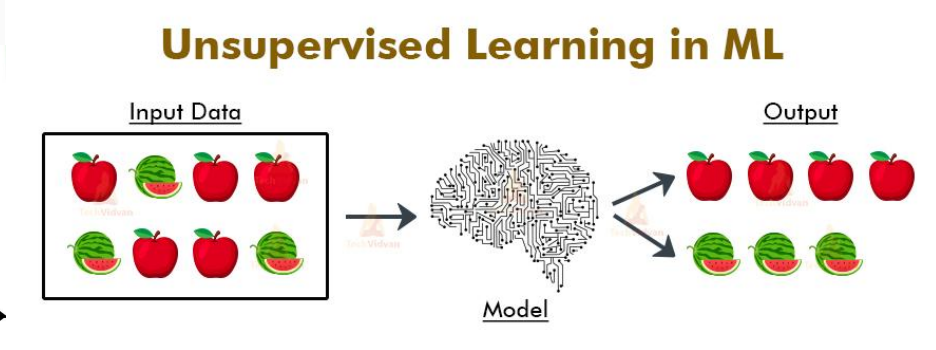




Types of “Learning”

- **Unsupervised Learning**
  - Algorithms trained using **unlabeled** (/unknown) data
  - There is no **feedback**
  - Can be categorized in **Clustering** and **Association**
  - Example: Fruit classifier



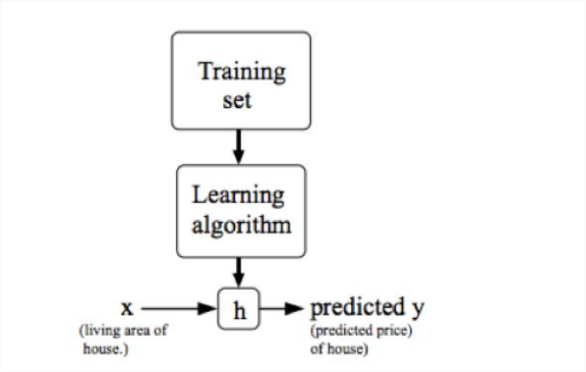


Supervised Learning

- Given a training set, learn a function `h(Hypothesis):X → Y` so that `h(x)` is a “good” predictor for the corresponding value of `y`

Regression: When the target variable is continuous - Housing Price Prediction

Classification: When the target variable can only take a small number of discrete values such as predicting if it’s a house or an apartment

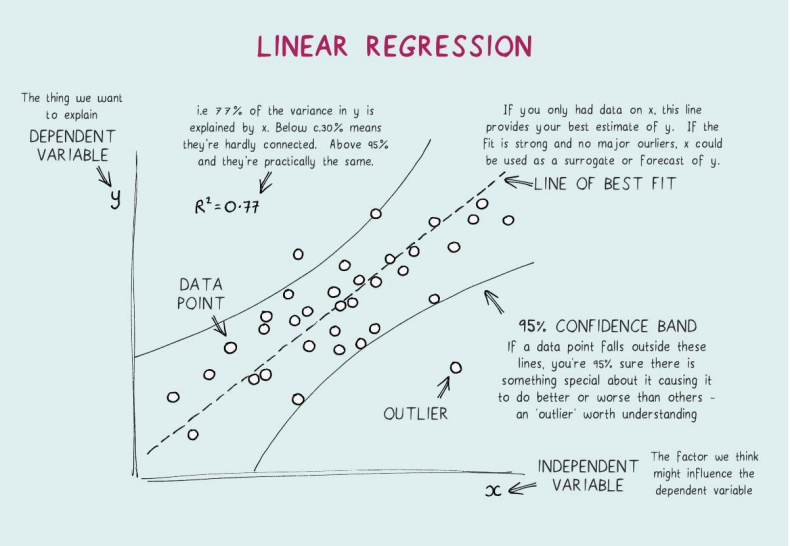


Linear Regression

- Model the relationship between two variables by fitting a **linear equation** to **observed data**
- A **continuous variable** being available
- **Estimating** the dependant

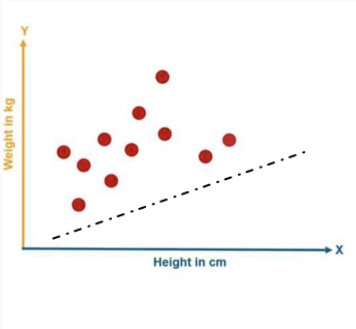
The general equation of a linear regression would be

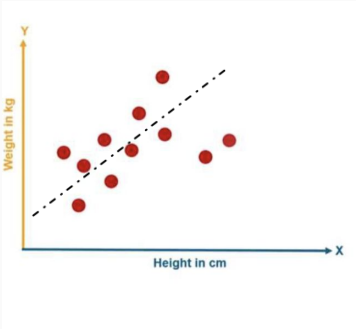
$$y = ax + b$$

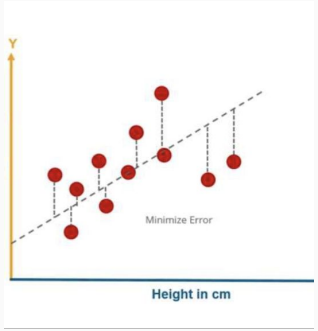


Example

- Start with a random line `w = h`
- Check error in prediction (hypothesis) and update
- Stop when no change or threshold is reached





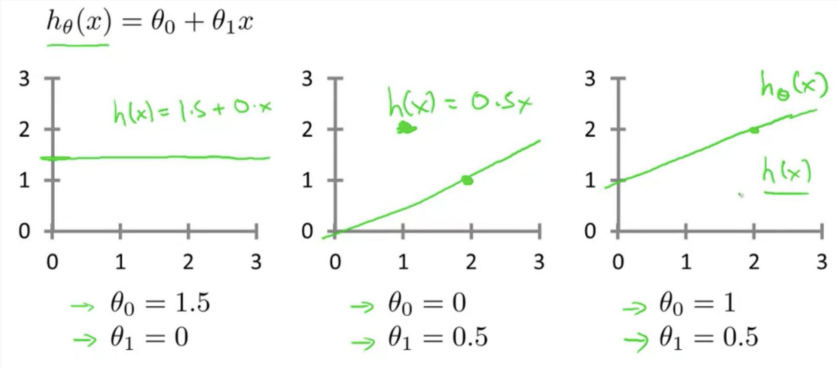


Hypothesis

Training Set

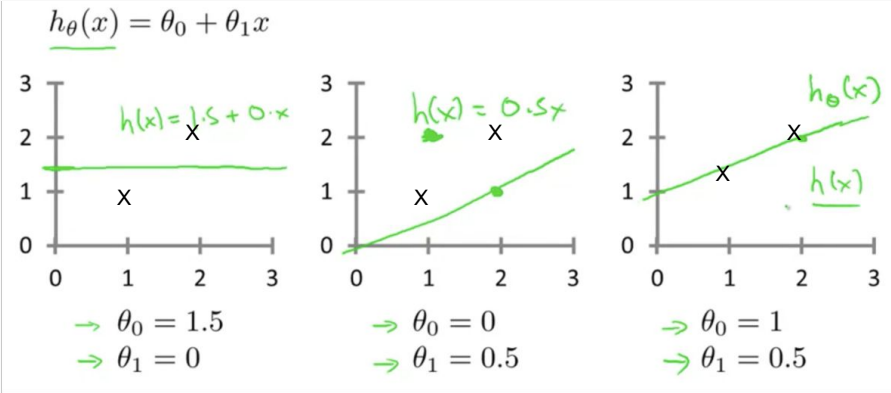
Size in feet² (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$



Cost Function

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$



Cost: $((1.5-1)^2 + (1.5-2)^2)/2*2$ = 0.5/4

Cost: ?

Cost: $((1-1)^2 + (2-2)^2)/2*2$ = 0

Cost Function

- Accuracy of hypothesis can be measured by using a **cost function**
- In the previous example, select the values such that the cost is minimum

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

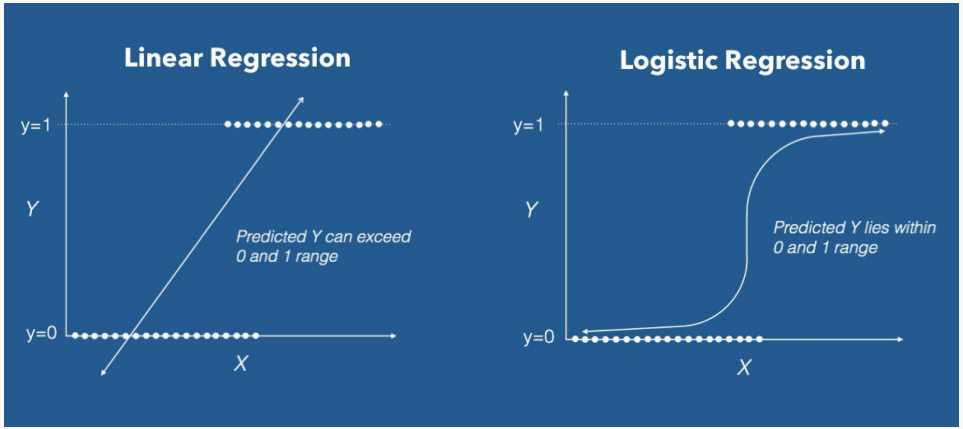
Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

- It is an **average of squared difference** between actual output (y) and the hypothesis results in the input (x)

Logistic Regression

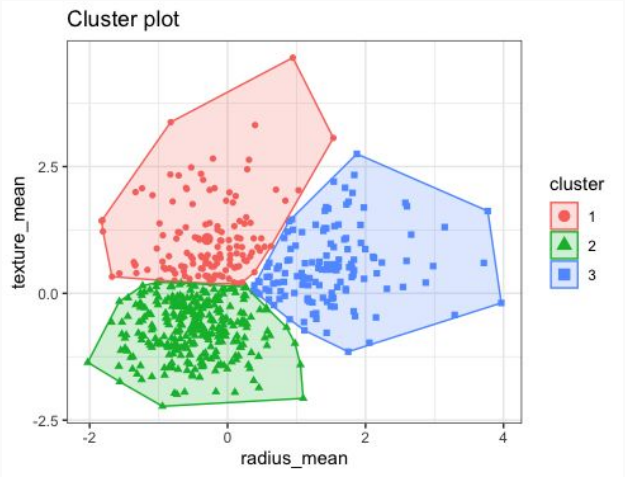
- Modeling the **probability** of a certain class or event
- Can be extended to model **several classes of events**
- Generally used for **classification**
- Has a categorical response

$$P(\text{Event}) = \frac{P(\text{Occurrence})}{P(\text{Not Occurrence})}$$



Clustering

- Automatically discovering **natural grouping** in data
- Unlike **supervised learning**, clustering only interprets the input data and find natural clusters in feature space



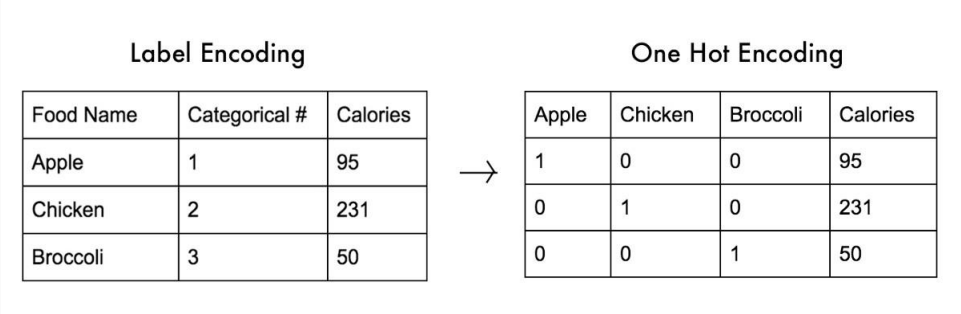
Model Representation

•	$x^{(i)}$	⇐ Input variables / features
•	$y^{(i)}$	⇐ Output variables / features
•	$(x^{(i)}, y^{(i)})$	⇐ Training Example
•	X/x	⇐ Space of input values
•	Y/y	⇐ Space of output values

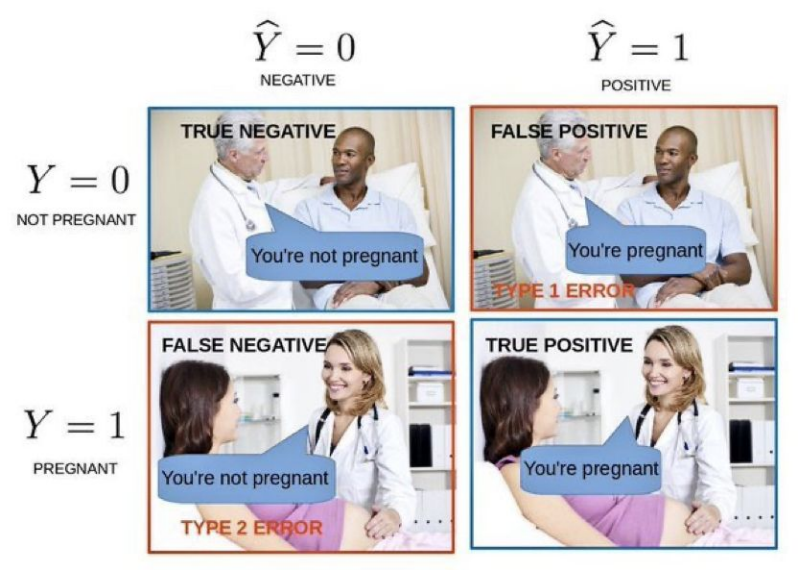
Encoding

- Converting categorical data into numerical data to fit and evaluate in the model

- There are multiple methods of encoding



Confusion Matrix



		Predicted class: P	Predicted class: N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Summary

- Supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not
  - Unsupervised learning models work on their own to discover the inherent structure of unlabeled data
- Confusion matrix provides accurate insight into how correctly the model has classified the classes depending upon the data fed or how the classes are misclassified

That’s all Folks!

See you in the next session :)

Give us a feedback: https://bit.ly/3g6ZDlD