GENERAL ASSEMBLY

Decision Trees & Random Forests

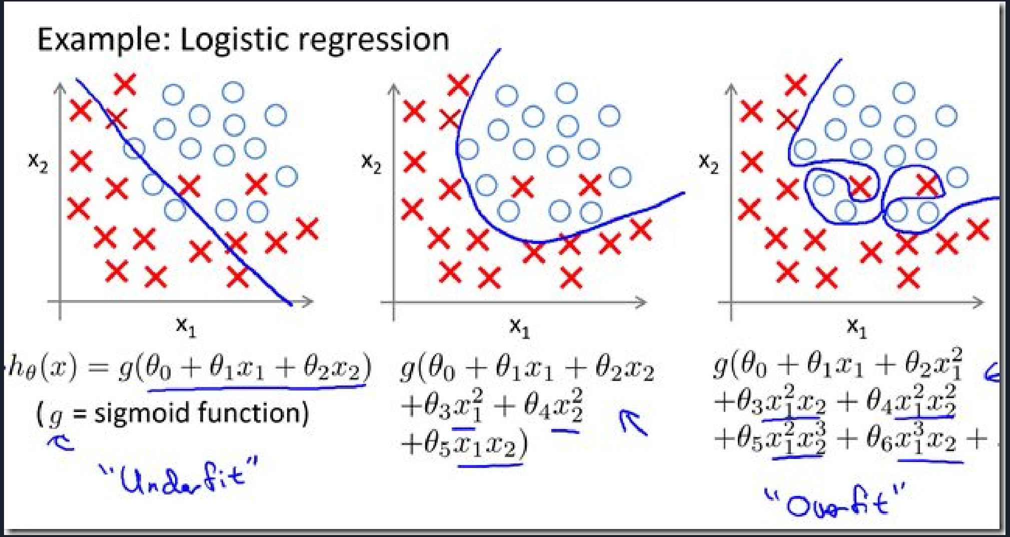
Data Science Curriculum

Last week we learnt..

- 3x Lessons - Logistic Regression for Classification tasks
  - Classifies 2 class problems, e.g. will respond to marketing campaign or will not respond

- Advantages?

- Disadvantages?



Example: Logistic regression

$h_\theta(x) = g(\underline{\theta_0 + \theta_1 x_1 + \theta_2 x_2})$

(g = sigmoid function)

"Underfit"

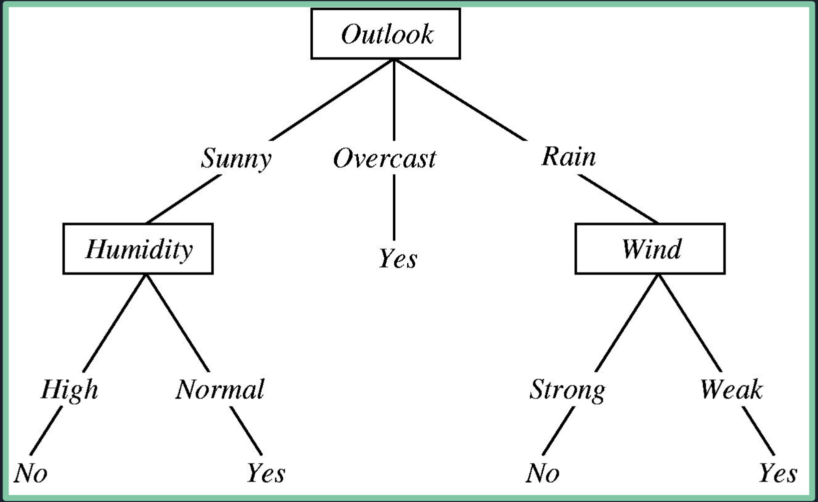
$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2$
$+\theta_3 x_1^2 + \theta_4 x_2^2$
$+\theta_5 \underline{x_1 x_2})$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2$
$+\theta_3 \underline{x_1^2 x_2} + \theta_4 \underline{x_1^2 x_2^2}$
$+\theta_5 \underline{x_1^2 x_2^3} + \theta_6 \underline{x_1^3 x_2} +$

"Over-fit"

Lesson Objectives

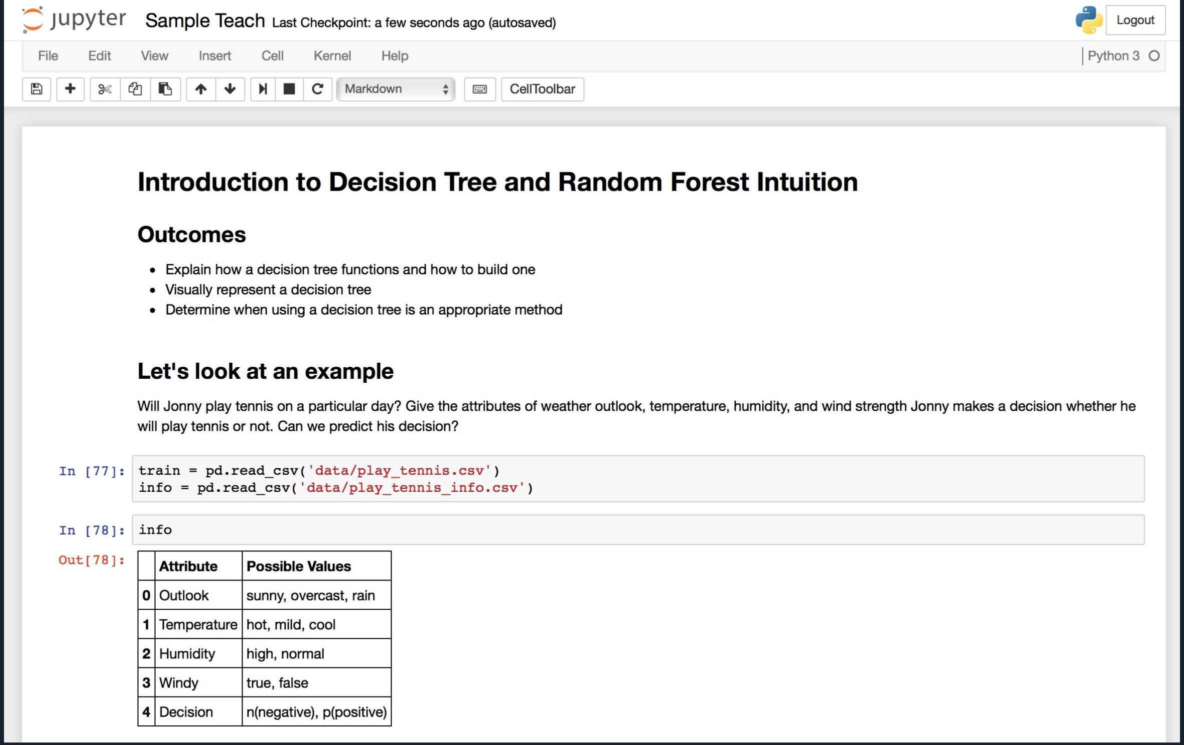
1. Understand how a decision tree functions and how to build one
2. Learn to visually represent a decision tree
3. Understand when to use a decision tree



Agenda

1. To Play or Not to Play…. Tennis Example of decision trees
2. Worked example of Iris Decision Trees in Python using Scikit-Learn
3. When to use decision trees
4. Intro to Random Forests

Jupyter Notebook



Introduction to Decision Tree and Random Forest Intuition

Outcomes

- Explain how a decision tree functions and how to build one
- Visually represent a decision tree
- Determine when using a decision tree is an appropriate method

Let's look at an example

Will Jonny play tennis on a particular day? Give the attributes of weather outlook, temperature, humidity, and wind strength Jonny makes a decision whether he will play tennis or not. Can we predict his decision?

```
In [77]: train = pd.read_csv('data/play_tennis.csv')
         info = pd.read_csv('data/play_tennis_info.csv')
```

```
In [78]: info
```

Out[78]:

	Attribute	Possible Values
0	Outlook	sunny, overcast, rain
1	Temperature	hot, mild, cool
2	Humidity	high, normal
3	Windy	true, false
4	Decision	n(negative), p(positive)

The End

By Andrew Szwec

Split data on column labels then calculate number of outcomes

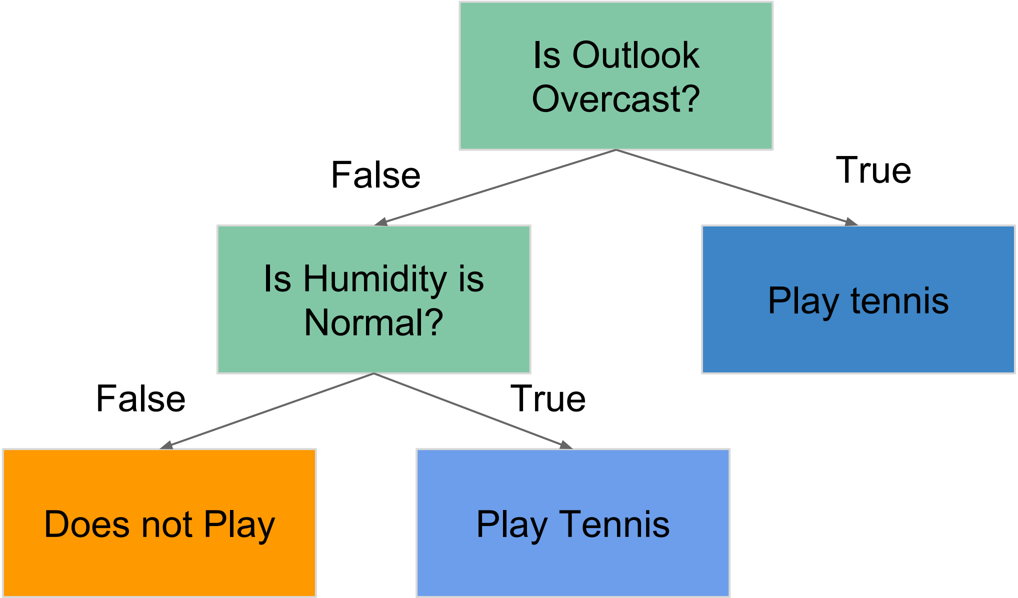
https://youtu.be/eKD5gxPPeY0

Decision Trees for Classification

- Decision trees for classification - classes like dog/cat, will purchase/won’t purchase
- Decision trees for regression - predicts a value like house price
- Make a decision tree for taxi dataset
- Visualise it
- Now use a toy dataset so they can do it themselves

Prune tree using validation set. Take away each node separately and see how performance improve/degrades against the validation set

Decision Tree



Is Outlook Overcast?
False
True
Is Humidity is Normal?
Play tennis
False
True
Does not Play
Play Tennis