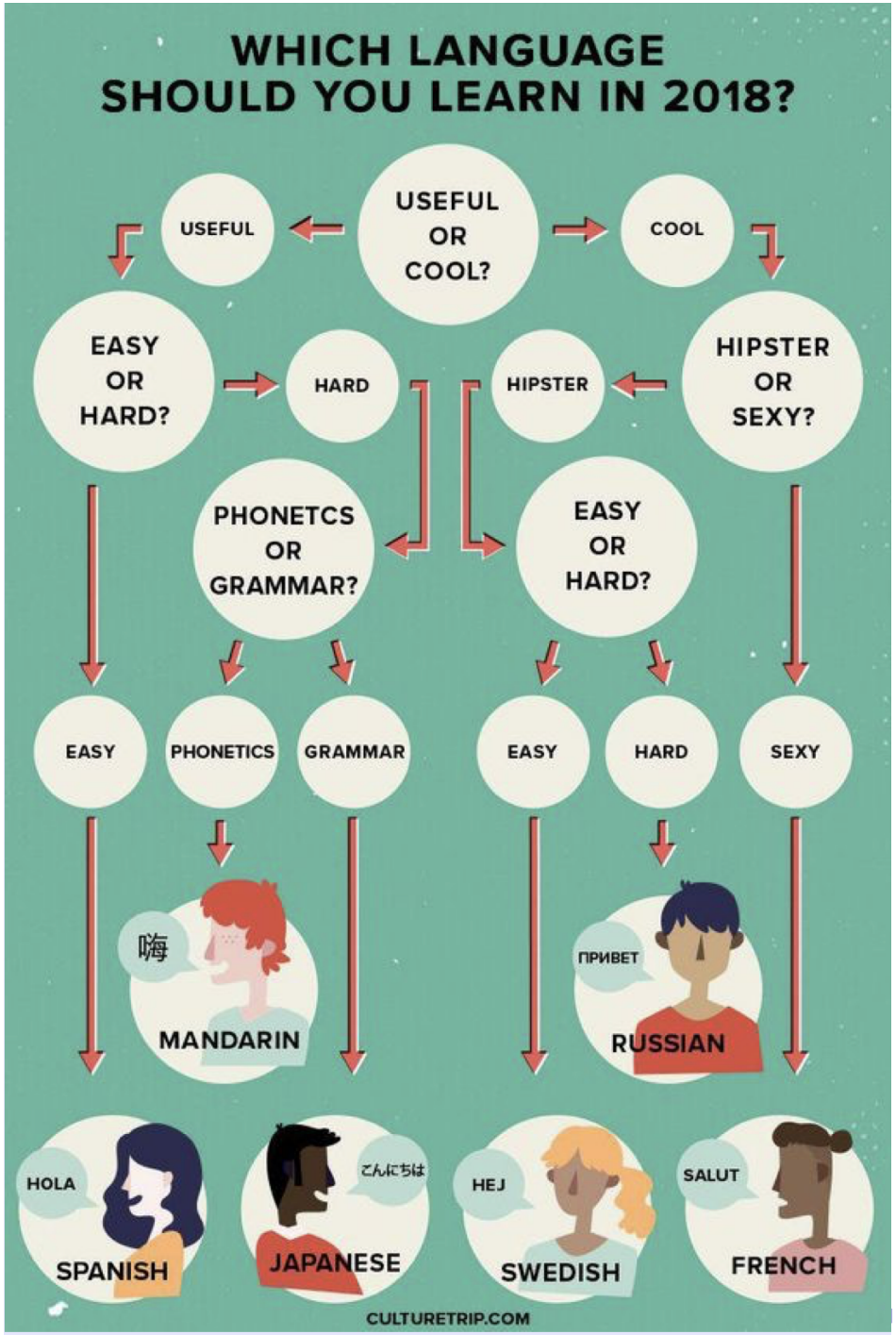
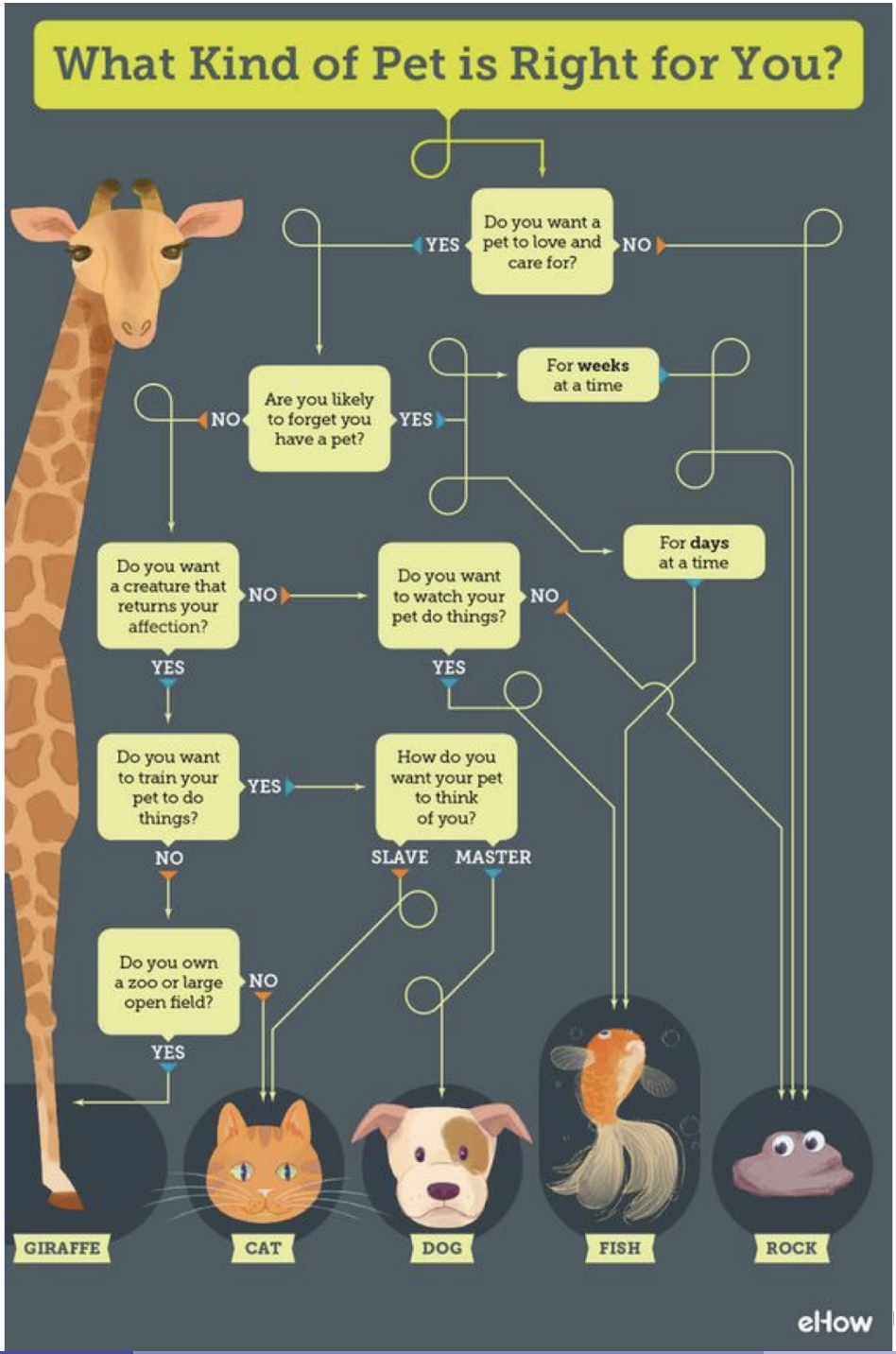
Decision Trees

Sujay Sanghavi


WHICH LANGUAGE SHOULD YOU LEARN IN 2018?
USEFUL OR COOL?
USEFUL
COOL
EASY OR HARD?
HARD
HIPSTER
HIPSTER OR SEXY?
PHONETCS OR GRAMMAR?
EASY OR HARD?
EASY
PHONETICS
GRAMMAR
EASY
HARD
SEXY
嗨
MANDARIN
ПРИВЕТ
RUSSIAN
HOLA
SPANISH
こんにちは
JAPANESE
HEJ
SWEDISH
SALUT
FRENCH
CULTURETRIP.COM


What Kind of Pet is Right for You?
Do you want a pet to love and care for?
YES
NO
Are you likely to forget you have a pet?
NO
YES
For weeks at a time
For days at a time
Do you want a creature that returns your affection?
NO
YES
Do you want to watch your pet do things?
NO
YES
Do you want to train your pet to do things?
YES
NO
How do you want your pet to think of you?
SLAVE
MASTER
Do you own a zoo or large open field?
NO
YES
GIRAFFE
CAT
DOG
FISH
ROCK
eHow

Tree Methods

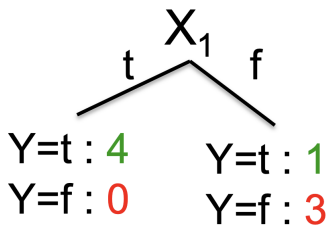
Idea: To get from x (the features) to y (the label / target to be predicted), do a **series of if-else questions** until you reach a decision (for classification) or prediction (for regression)

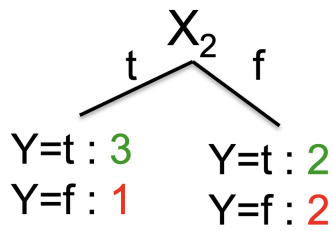
Tree Methods

Idea: To get from x (the features) to y (the label / target to be predicted), do a **series of if-else questions** until you reach a decision (for classification) or prediction (for regression)

Step 1: Choosing which feature to split on.

X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F
F	T	F
F	F	F



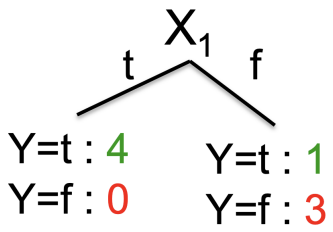


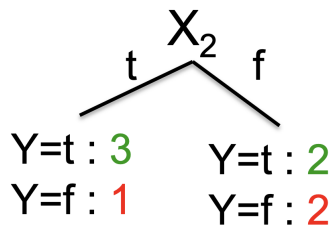
Tree Methods

Idea: To get from x (the features) to y (the label / target to be predicted), do a **series of if-else questions** until you reach a decision (for classification) or prediction (for regression)

Step 1: Choosing which feature to split on.

X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F
F	T	F
F	F	F





Choose the feature that leaves us the most certain after splitting on it.

Measuring Uncertainty

Which of the following two distributions is more uncertain ?

P(Y=A) = 1/2	P(Y=B) = 1/4	P(Y=C) = 1/8	P(Y=D) = 1/8

P(Y=A) = 1/4	P(Y=B) = 1/4	P(Y=C) = 1/4	P(Y=D) = 1/4

Tree Methods

Basic idea: First partition the training data samples. Then, to find prediction $\widehat{y}(x)$ for a query x:

- Find the partition that it goes into.
- Then

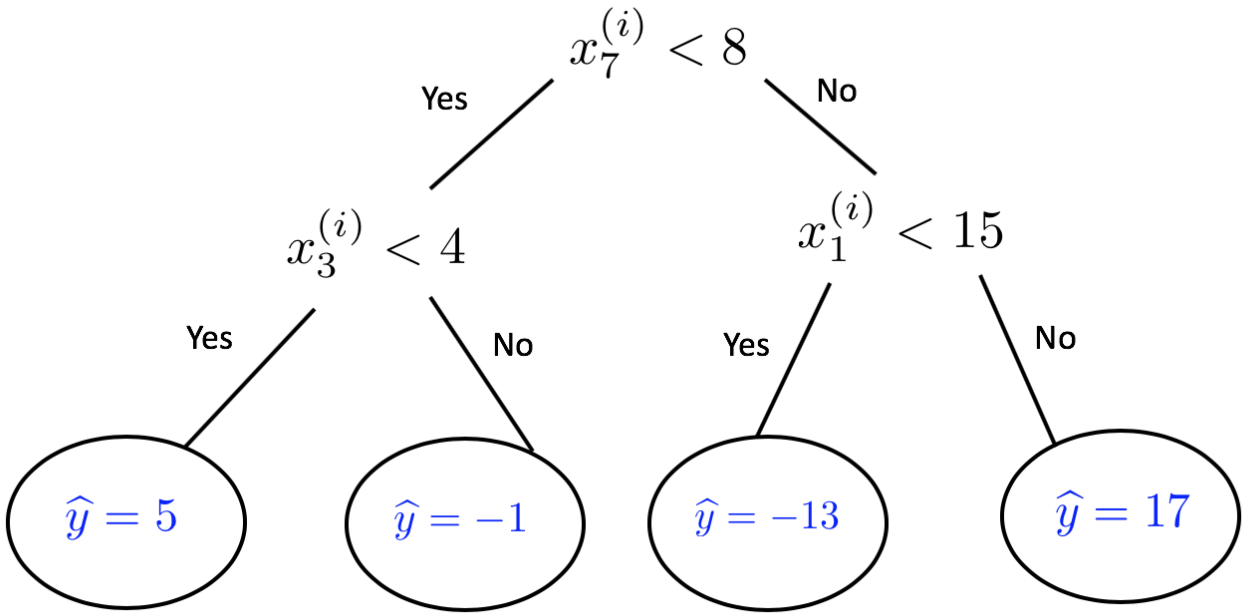
$$\begin{aligned} \widehat{y}(x) &= \text{average of all training sample } y\text{'s in that partition} \\ &\quad \text{(for regression)} \\ &= \text{majority of all training sample } y\text{'s in that partition} \\ &\quad \text{(for classification)} \end{aligned}$$

Tree models involve recursively partitioning the samples by

- Choose a feature x_j and a threshold τ.
- For each sample $x^{(i)}$, put it in the "left bucket" if $x_j^{(i)} < \tau$ and "right bucket" otherwise

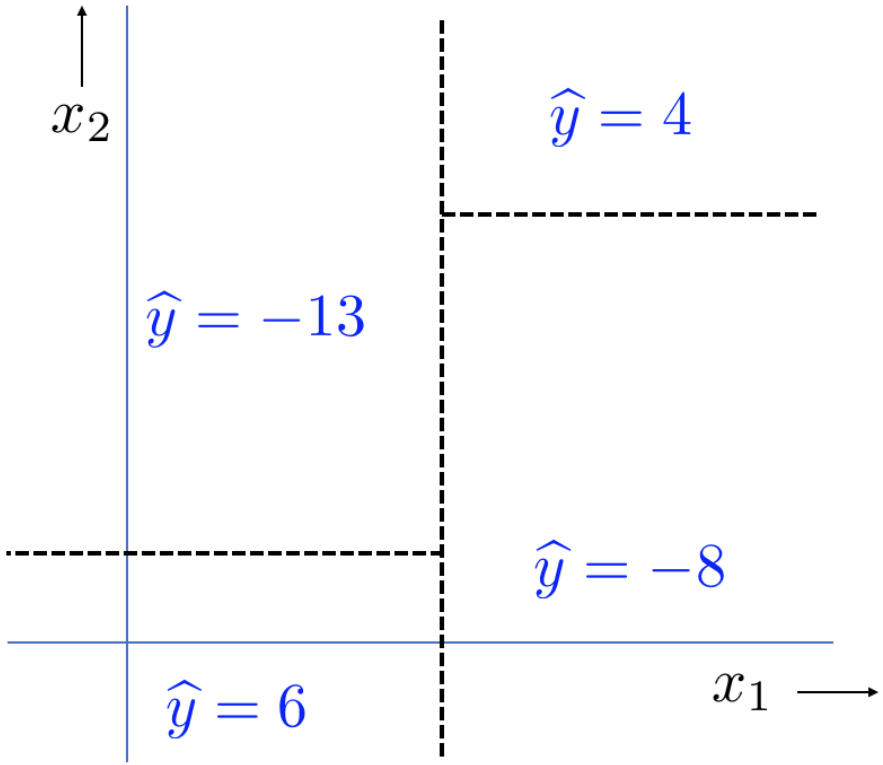
Decision / Regression Trees

Decision rule represented by a tree



Decision / Regression Trees

Decision regions can be visualized as a series of "stumps"

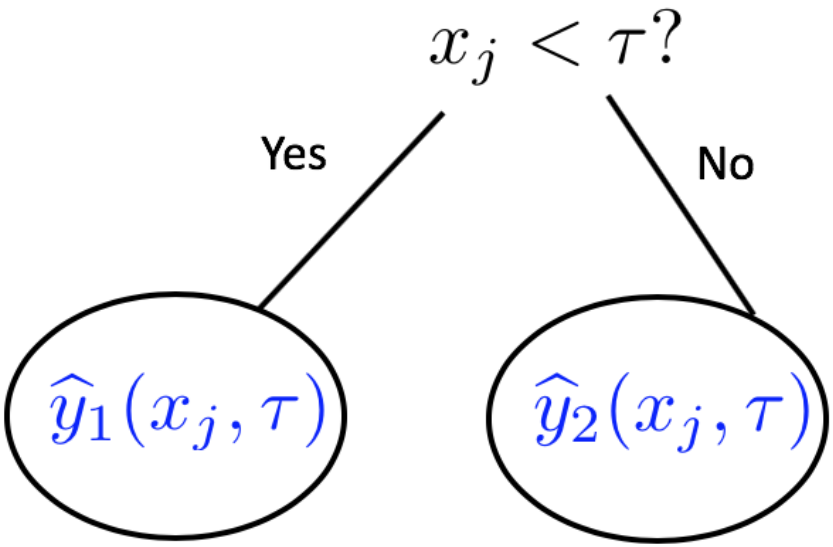


Decision / Regression Trees

Q: How to choose the feature and thresholds for each split ?

A1: Greedy: at each step, choose the feature and threshold that gives the lowest final squared error.
If feature x_j and threshold τ, final squared error:

$$\sum_{i \in S_1(x_j,\tau)} \left(y^{(i)} - \widehat{y}_1(x_j,\tau)\right)^2 + \sum_{i \in S_2(x_j,\tau)} \left(y^{(i)} - \widehat{y}_2(x_j,\tau)\right)^2$$



However, for a large number of features, this is very cumbersome.

Also: bigger depth == lower training error, but more likely to overfit ...