Introduction to Machine Learning and DecisionTree

What is machine Learning?

* Inducing a rule (model) from past data and outcomes.

Example 1:
Tourists came to the USA for the first time and knowing nothing about the USA observed after a few weeks all offices are open on weekdays but closed on the weekends.

Data (Observations) :
Day is on weekend. Office is closed.
Day is on weekday. Office is open
Induction (Rule):
Offices are closed on weekends(Sat/Sun) in the USA.

Example 2:
Pavlov's dog was a machine learning model

Data (Observations) :
observe Bell => get Food
observe no Bell => get no food
Induction (Rule)
Bell => Food

Example 3:

X	0	2	3	4	5	6	7	8
Y	0	4	6	8	10	12	14	16

Learn a rule

y = 2x (known as linear regression)

Now we can deduce Y from never before seen X. Need to only know rule. Not data anymore.

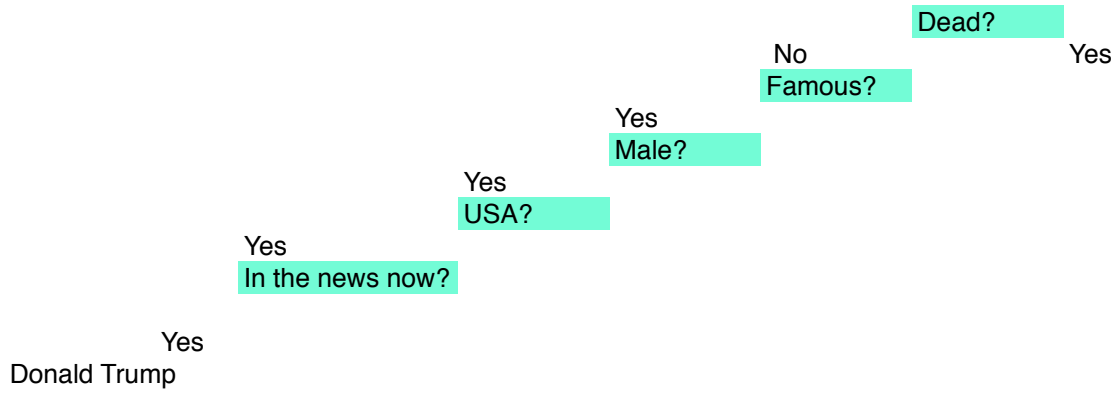
Goal of Machine Learning

* Once a model (rule is learned) predict outcomes when new unseen data is seen.

DecisionTree

* One of the machine learning models. Can be used for both regression/classification.
* Exactly the way humans process information to achieve information gain

* Play 20 questions game (famous person)



In Machine Learning given:

Data:
    Features: Dead, Famous, Sex, Country, News
    Label: Donald Trump

Learn:
    The tree we just drew above.

Can we predict Michael Jackson from this Tree?

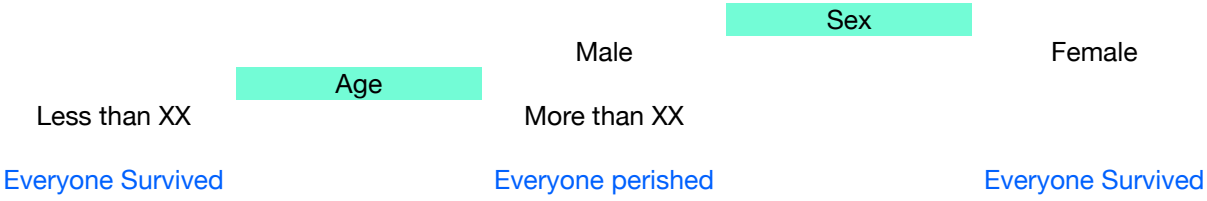
* Importance of question asked at every step (feature selection based on Information gain)
    => Was our first question good?
    => What if our first question was: Does the person like broccoli?
* Importance of limiting number of questions (prevent overfitting)
    => How would the game work if allowed to ask a million questions instead of 20
* Importance of having adequate data for robust model creation (prevent bias)
    => What if we did not know answer to most questions asked? Can we create a tree?
* http://www.20q.net/ (found this on the web. Uses some AI to learn based on what users told it)

Titanic Dataset

Everyone perished. (`Predictions have an accuracy of 60.00%.`)

First Feature Selection: Sex (`Predictions have an accuracy of 78.68%.`)

Second Feature Selection: Age (`Predictions have an accuracy of 80.02%.`)



XX: what is XX? I know what it is. You need to find out.

Point is: Prediction accuracy increased (entropy reduced, information gained) as features selected. We were making "Educated" guesses instead of "Wild" guesses.

Question: Are Decision trees unique? Is this the only DT that can give good predictions?

Question: What is the use of this model? What question can it answer that we couldn't answer some other way?

Homework (Next Class Discussion)

1) Read up on equation of information gain and entropy. Does the equation make sense related to how we used it in the DecisionTree? Why do you think it is a good measure to use for feature selection?

2) Read up on RandomForest. What do you think they are? How are they different than DecisionTree? What are the advantages and disadvantages of RandomForest?