Decision Trees

Session# 2

Introduction

- One of the most simplest and interpretable classifiers/regressors in the family of models for analytics
- Supervised learning algorithm. It requires data to be labeled.
- Non parametric learning algorithm
- Cut space into disjoint subsets as decision boundaries in the feature space

Introduction (Cont.)

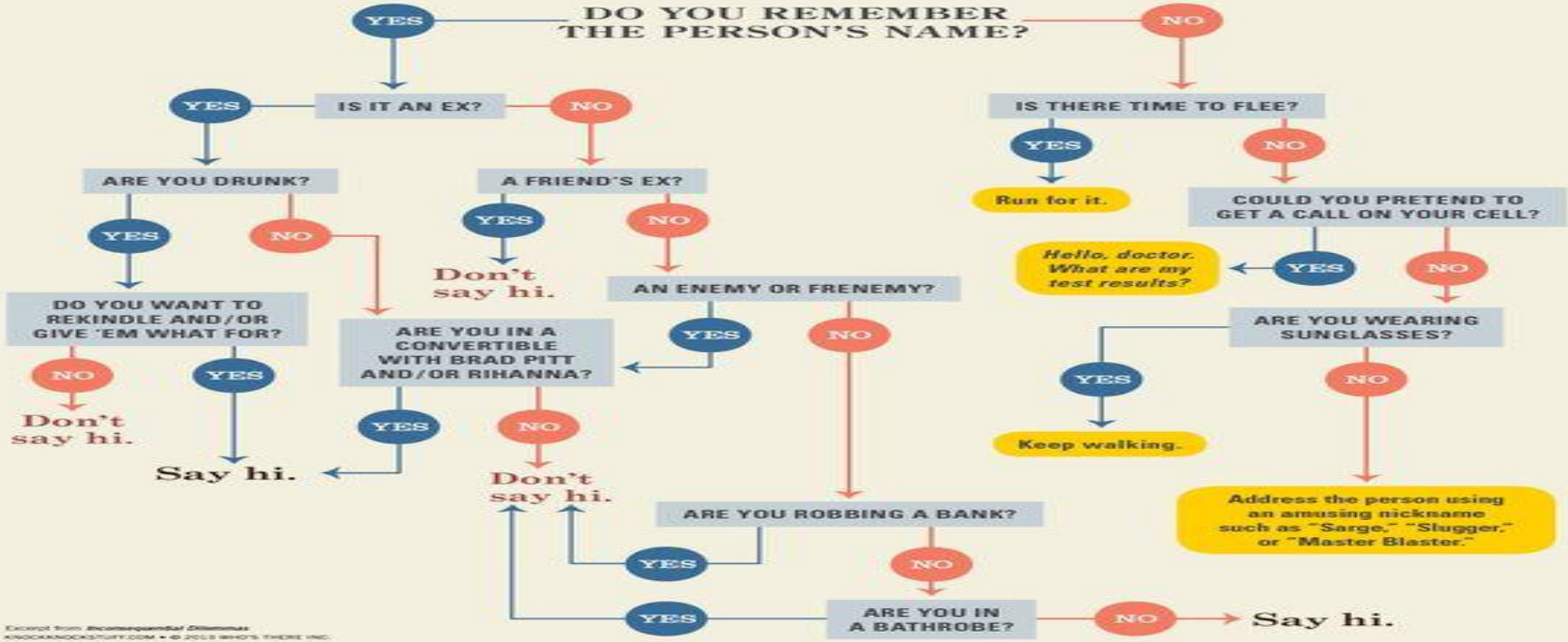
- Decision trees learn **rules** from data (similar to rule-based engines in classical AI, but rules are learned instead of hard coded)

- Tree ask questions relevant to classification/regression at each step in tree.

- Let’s see an example of the tree

I just saw someone I know.

DO I SAY HI?

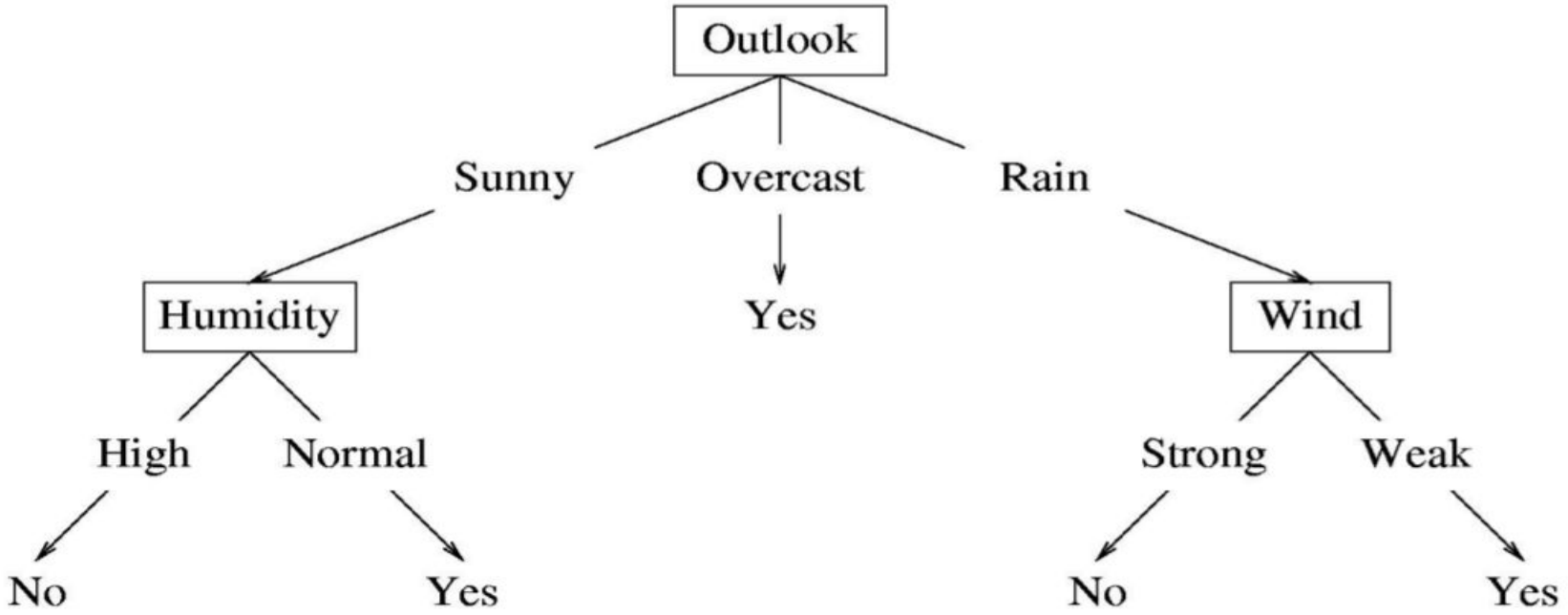


Example Data

$\langle \boldsymbol{x}_i, y_i \rangle$

Predictors				Response
Outlook	**Temperature**	**Humidity**	**Wind**	**Class**
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Tree Learned



Terminologies related to Tree

- Node : each node in the tree is associated with any input feature/predictor/variable. Root node is the top-most node of the tree.

- Branches associated with each input is related to the type of the variable:
  - **Categorical variable:** every branch correspond with one class/category
  - **Numerical variable:** usually two branches based on a threshold

- Leaf Node : Last level of the tree. This is where prediction is performed for the samples/instances/example.

Building Tree

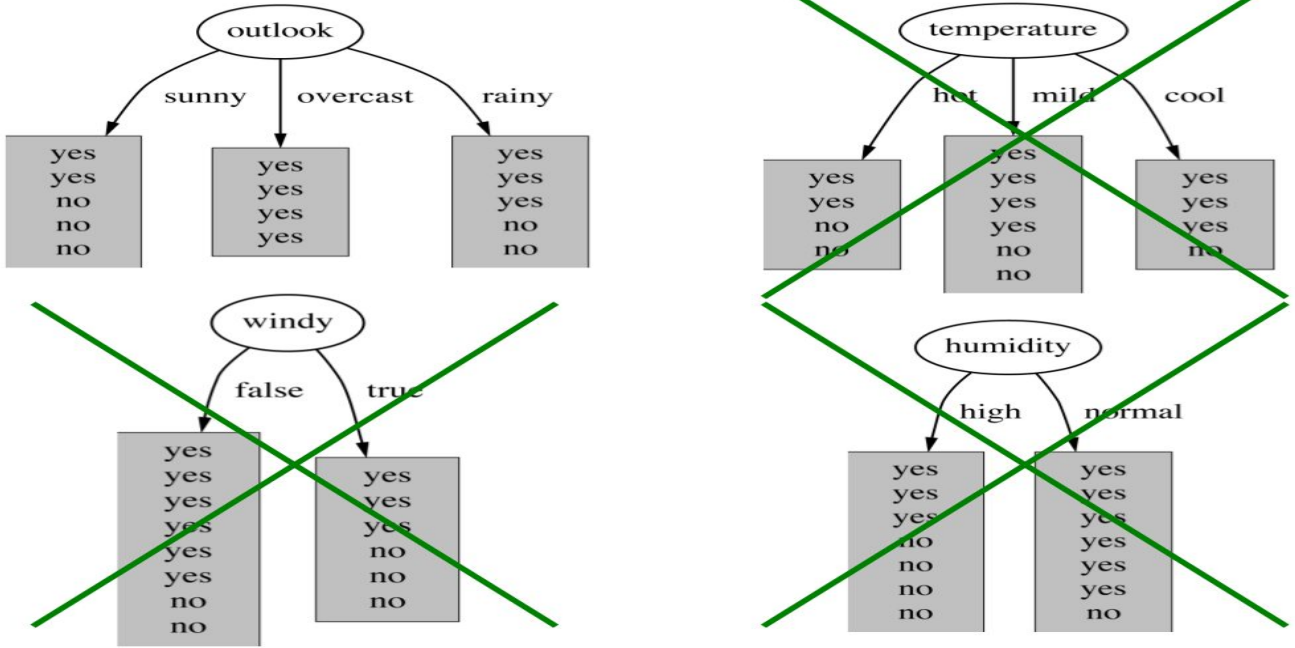
- Growing a tree boils down to select attributes/variables/features for each level of tree.

- We need to place attributes at each level in a way which results in higher accuracy on testing set. (**Generalizability is key to Machine learning!**)

- Consider the data again in the next slide

Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
07-30	mild	rain	high	false	yes

Selecting attribute for root node

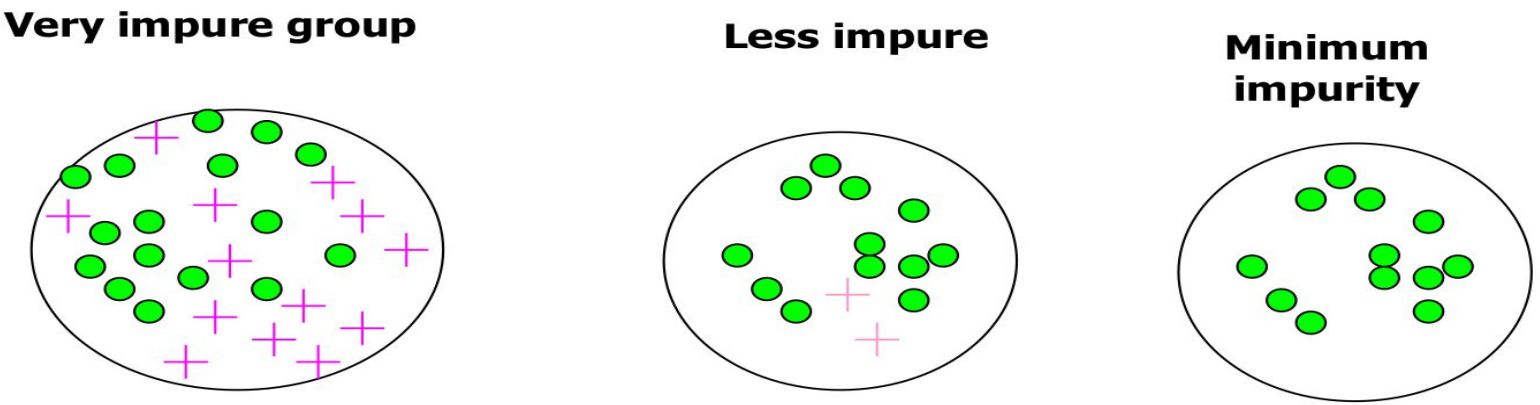


What is a good Attribute?*

- A good tree prefers attributes that split the data so that each successor node is as **pure** as possible,i.e. the distribution of examples in each node is so that it mostly contains examples of a single class
- **Idea**: A good attribute splits the examples into subsets that are (ideally) “ all positive” or “all negative”
- Entropy is used to quantify the purity of a node. It is a concept taken from Information theory. **Lesser the entropy, the more pure is the node and vice versa.** Technically, it measures impurity instead of purity.
- In essence, we will select the attribute resulting in **lowest entropy**.

A word about Entropy

- Entropy = $\sum -p_i \log_2 p_i$
- P_i is the probability of class i. It is computed as the proportion of class i in the set.



Back to growing trees

- At each step of tree, we would select the attribute which result in more pure node.

- The choice is greedy. We don’t consider long run effects of selection of node.

- For classification, selecting attribute which minimize **entropy**

  For regression, selecting attribute which minimize **sum of squared deviation from mean**

Doing predictions in tree

- Say you have trained your model on the data (in other words, you build your tree from the data). Trying to be fancier is healthy and insane at the same time. Choose wisely :)

- To predict for any new sample, propagate the sample towards the leaf. Once you are at the corresponding leave, there are 3 possibilities:
  - Leaf has a single training instance. Assign the same label as of the training sample
  - Leaf has many training samples. Assign the label of majority class samples in that leaf
  - Leaf has many training samples. Assign the mean of samples in that leaf (**Regression)**

Modeling the data

Predicting Earnings from Census Data

- The US periodically collects demographic information by conducting a census

- Data contains 1994 census data for 31,978 individuals in US

- In the problem, we have to **predict** how much a person earns using the census information. To be exact, we need to predict if the person earns $50,000 per year or not.

Data Variables

- *age* = the age of the individual in years
- *workclass* = the classification of the individual's working status (does the person work for the federal government, work for the local government, work without pay, and so on)
- *education* = the level of education of the individual (e.g., 5th-6th grade, high school graduate, PhD, so on)
- *maritalstatus* = the marital status of the individual
- *occupation* = the type of work the individual does (e.g., administrative/clerical work, farming/fishing, sales and so on)
- *relationship* = relationship of individual to his/her household
- *race* = the individual's race
- *sex* = the individual's sex
- *capitalgain* = the capital gains of the individual in 1994 (from selling an asset such as a stock or bond for more than the original purchase price)
- *capitalloss* = the capital losses of the individual in 1994 (from selling an asset such as a stock or bond for less than the original purchase price)
- *hoursperweek* = the number of hours the individual works per week
- *nativecountry* = the native country of the individual
- *over50k* = whether or not the individual earned more than $50,000 in 1994

Let’s build a decision tree in R

Trees in R

- There are many available packages in R to construct trees from data.
- We will be using **rpart** package.

Evaluating the Model

- Precision: What proportion of positive identifications was actually correct?
- Recall: What proportion of actual positives was identified correctly?

- Need of precision and recall, when we have a metric such as accuracy ?