Classification Supervised Learning (Part1)

講者：Isaac

Outline

- Logistic Regression
- K-Nearest Neighbor
- Decision Tree
  - CART
  - ID3

Logistic regression

Logistic regression

$$Model\colon\ h_\theta = \frac{1}{1+e^{-\theta^T X}}\ where\ \theta = \begin{bmatrix}\theta_0\\ \vdots \\ \theta_n\end{bmatrix}, X = \begin{bmatrix}x_0\\ \vdots \\ x_n\end{bmatrix}$$

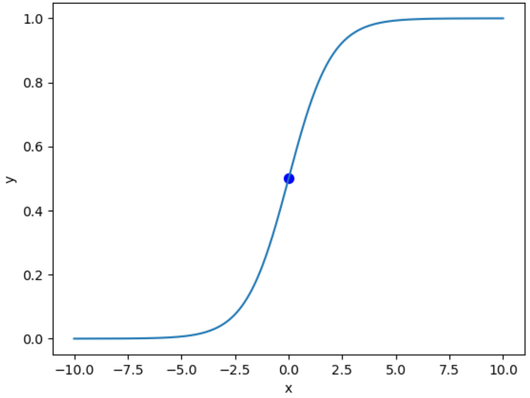
$$Learned\ Parameters\colon\ \theta_0, \theta_1, \dots, \theta_n$$

$$Cost\ Function\colon C(\theta_0, \theta_1, \dots, \theta_n)$$

$$= \frac{1}{2m}\sum_{i=1}^{m} y^{(i)}\log(h^\theta(x^{(i)})) + (1 - y^{(i)})\log(1 - h^\theta(x^{(i)}))$$

Logistic regression

- Sigmoid function
  - Output is [0, 1]



$$y = \frac{1}{1 + e^{-x}}$$

Sigmoid function

Logistic regression

- Actually, cost function in logistic regression is cross-entropy
  - note that cross-entropy can be used when each of output is probability distribution

$$\mathbf{p} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad H(\mathrm{p}, \mathrm{q}) = -\sum_{i} p_i \ln(\mathrm{q}_i) \qquad \mathbf{q} = \begin{bmatrix} 0.1 \\ 0.5 \\ 0.4 \end{bmatrix}$$

Logistic regression

- Information
  - $log\left(\frac{1}{p_i}\right)$ *where* p_i *is probability of an event*

Sun rises in the east tomorrow

It will rain tomorrow in Taiwan

Which is more informative?

Entropy V.S. Cross-entropy

- Entropy
  - Expected value(mean) of information contained in each message
- Entropy can be seen as index of uncertainty
  - Bigger mean more chaos
- Cross-entropy
  - Measurement on the difference between two probability distribution
  - Different distribution apply on entropy
  - Cross-entropy is greater than entropy

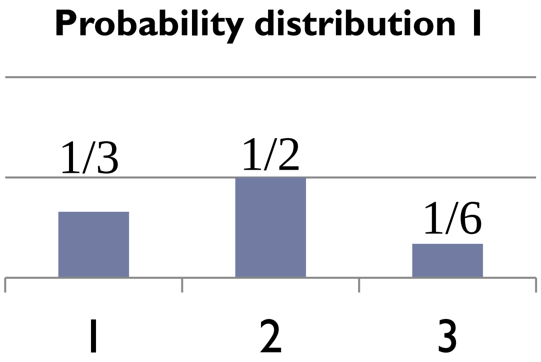
$$H(y) = \sum_i y_i \log\left(\frac{1}{y_i}\right) = -\sum_i y_i \log(y_i)$$

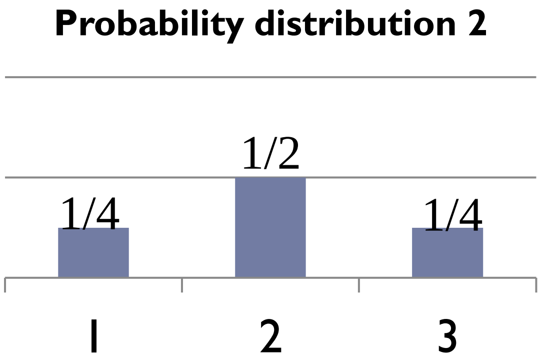
Entropy

$$H(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i)$$

Cross-entropy

Example





Entropy on distribution 1
= 1/3 * log(3) + 1/2 * log(2) + 1/6 * log(6)

Entropy on distribution 2
= 1/4 * log(4) + 1/2 * log(2) + 1/4 * log(4)

Cross-entropy on distribution 1 over distribution 2
= 1/3 * log(4) + 1/2 * log(2) + 1/6 * log(4)

Cross-entropy on distribution 2 over distribution 1
= 1/4 * log(3) + 1/2 * log(2) + 1/4 * log(6)

Example

Entropy on distribution 1
= 1/3 * log(3) + 1/2 * log(2) + 1/6 * log(6)
= 0.439

Entropy on distribution 2
= 1/4 * log(4) + 1/2 * log(2) + 1/4 * log(4)
=0.452

Cross-entropy on distribution 1 over distribution 2
= 1/3 * log(4) + 1/2 * log(2) + 1/6 * log(4) = 0.456

Cross-entropy on distribution 2 over distribution 1
= 1/4 * log(3) + 1/2 * log(2) + 1/4 * log(6) = 0.464

- Cross-entropy is greater than entropy
  Cross-entropy on distribution 1 over 2 > Entropy on distribution 1
  Cross-entropy on distribution 2 over 1 > Entropy on distribution 2
- If two distribution become closer
  - Value of cross-entropy is closer to entropy

Logistic regression

- Learning in logistic regression
  - Use gradient descent(same as linear regression)

Example and Practice

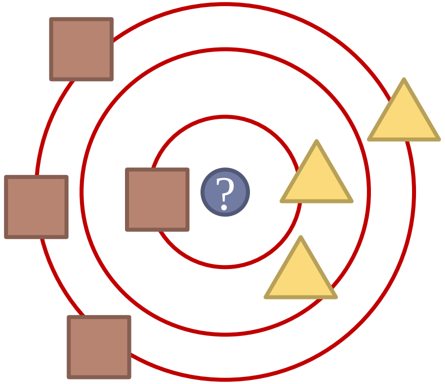
- Example
  - Logistic regression
    - example/regression
- Practice
  - Try to use linear regression to predict diabetes patient
    - dataset/pima-indians-diabetes.csv
    - practice/regression

  - More information about the dataset
    - https://www.kaggle.com/uciml/pima-indians-diabetes-database/data

K-Nearest Neighbor

What's K-Nearest Neighbor

- A non-parametric method used for classification and regression
- Also called kNN
  - "k" mean how many neighbors should be considered to help classification/regression

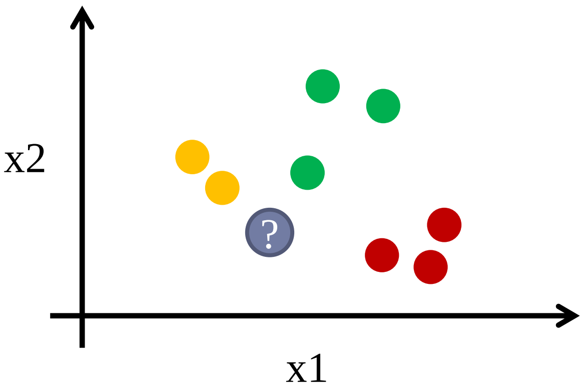


- k=1:
  - Belongs to square class
- k=3
  - Belongs to triangle class
- k=7
  - Belongs to square class

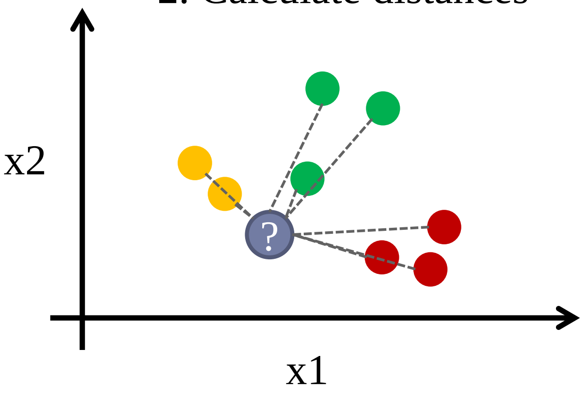
kNN intuitive concept

K-Nearest Neighbor

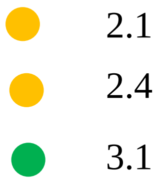
1. Look at the data



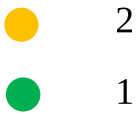
2. Calculate distances



3. Find neighbors



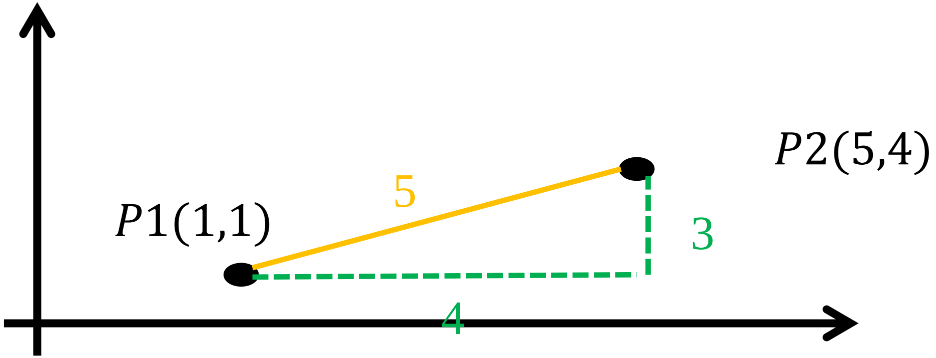
4. vote from labels



guess it is yellow class

How to Define Distance

- L1 distance (Manhattan distance)
- L2 distance (Euclidean distance)

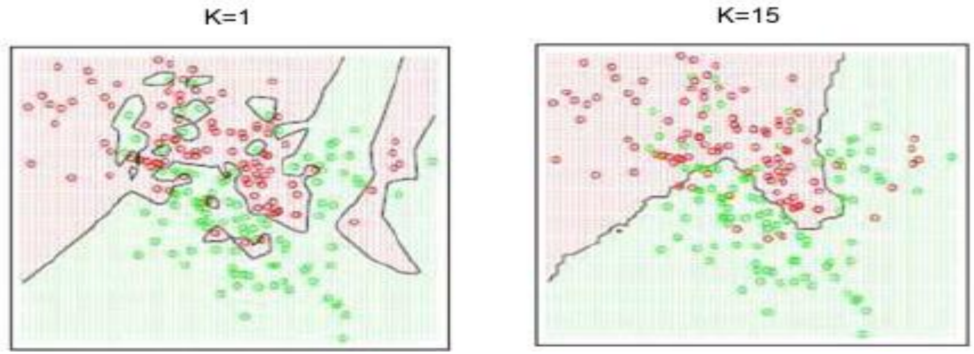


$$Euclidean\ distance = \sqrt{(5-1)^2+(4-1)^2} = 5$$

$$Manhattan\ distance = |5-1| + |4-1| = 7$$

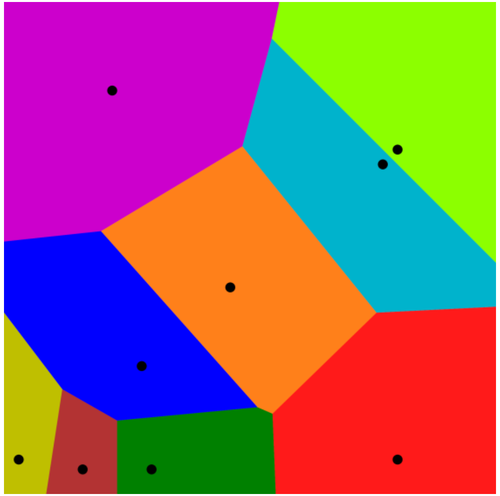
How to choose K?

- K is small
  - sensitive to noise points
- K is large
  - neighborhood may include points from other classes
  - smoother boundary
  - If too large, machine always predict majority class



- http://vision.stanford.edu/teaching/cs231n-demos/knn/

- 1-NN
  - Voronoi Diagram



Problem in L2 Distance

0 1 1 1 1

1 1 1 1 0

vs

1 0 0 0 0

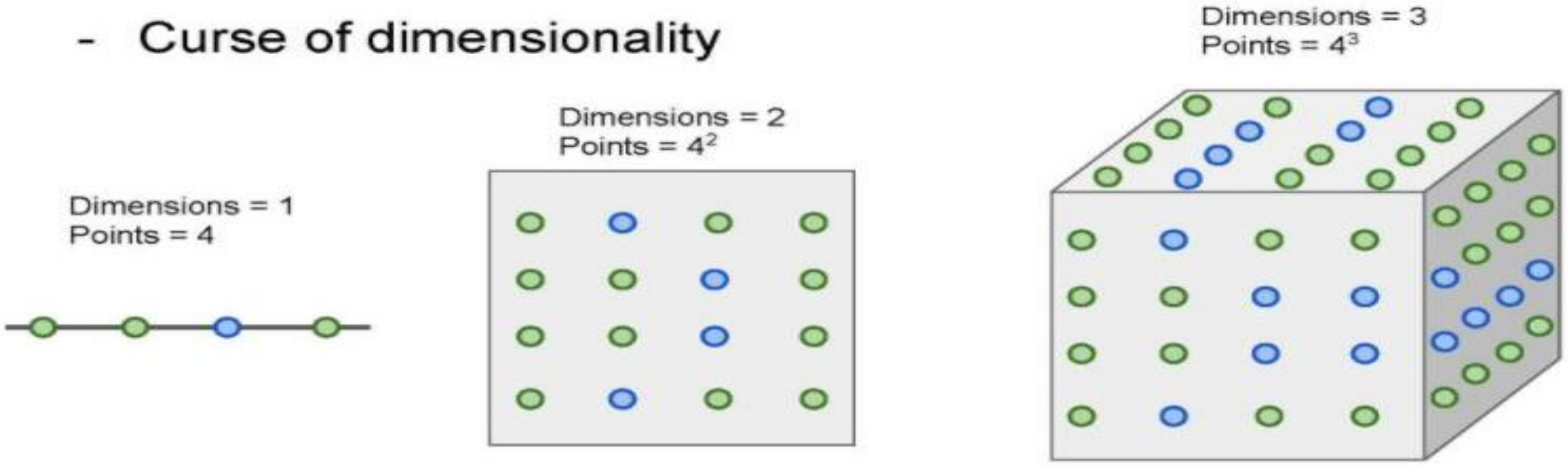
0 0 0 0 1

distance = 1.41

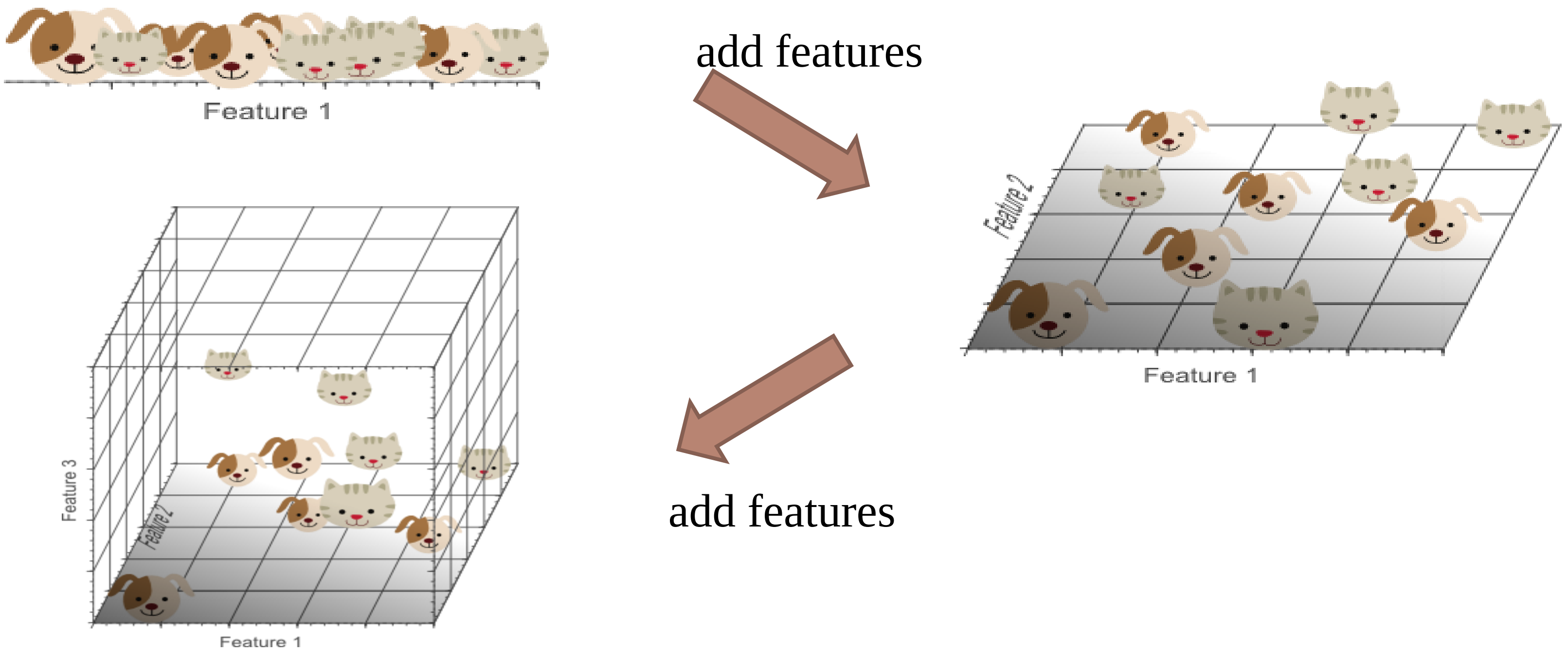
distance = 1.41

counter-intuitive results

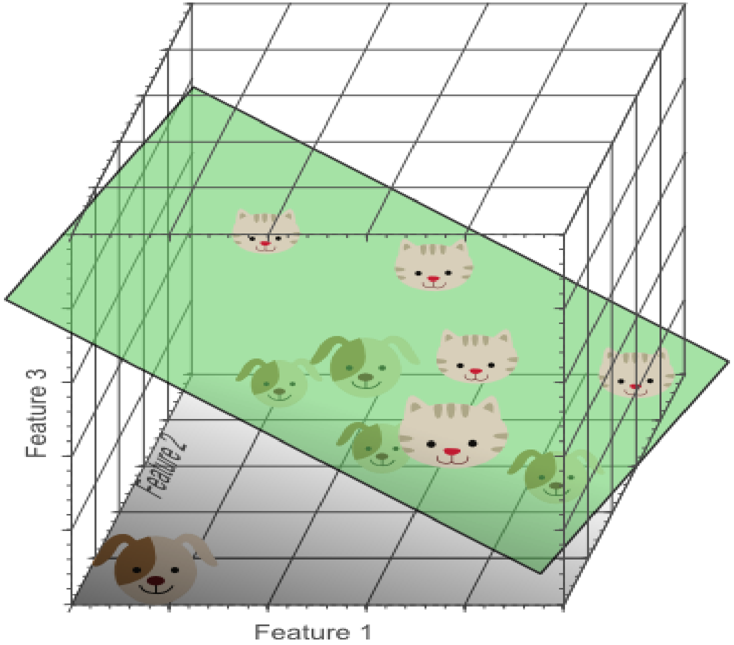
- Curse of dimensionality



Curse of dimensionality



Curse of dimensionality



Linear separable in high dimensionality

Curse of dimensionality

- Increase dimensionality may obtain perfect classfication
- However, extend too many dimensionality(features) lead to overfitting

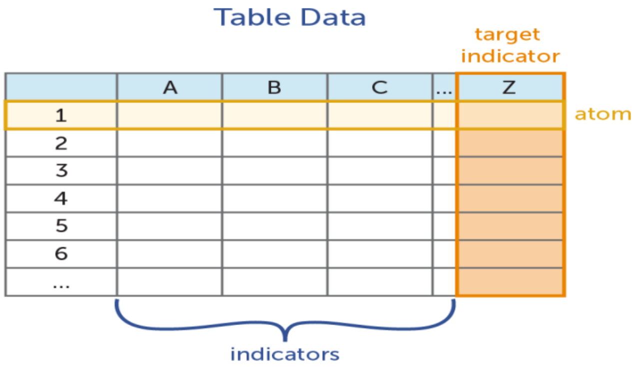
Example and Practice

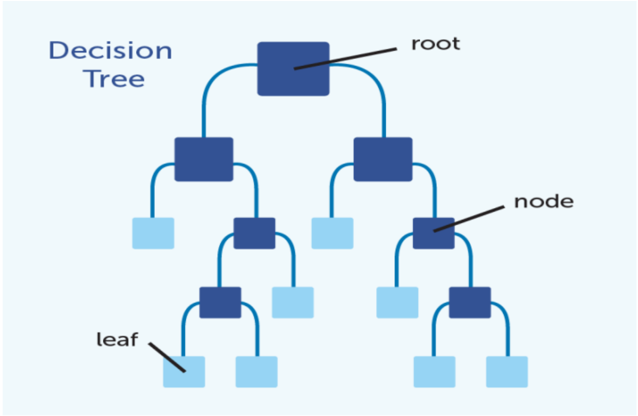
- Example
  - KNN
    - example/supervised learning
- Practice
  - Try to use knn to predict different varieties of wheat
    - dataset/seeds_dataset.csv
    - practice/supervised learning
  - More information about the dataset
    - https://archive.ics.uci.edu/ml/datasets/seeds#

Decision Tree

What's Decision Tree

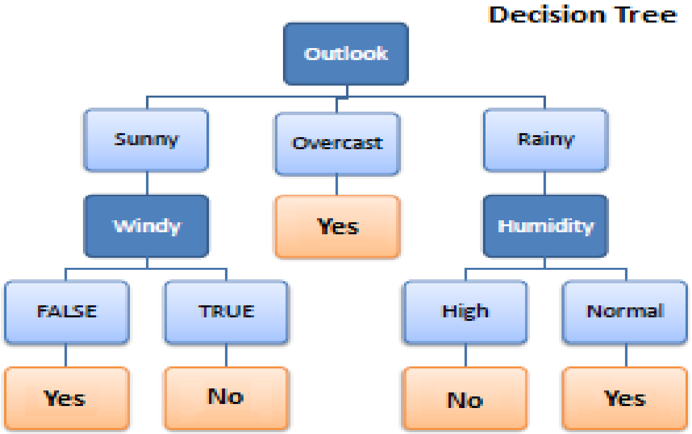
- A decision support tool that uses a tree-like graph of decisions and their possible consequences
- Common method in decision tree
  - ID3
  - CART



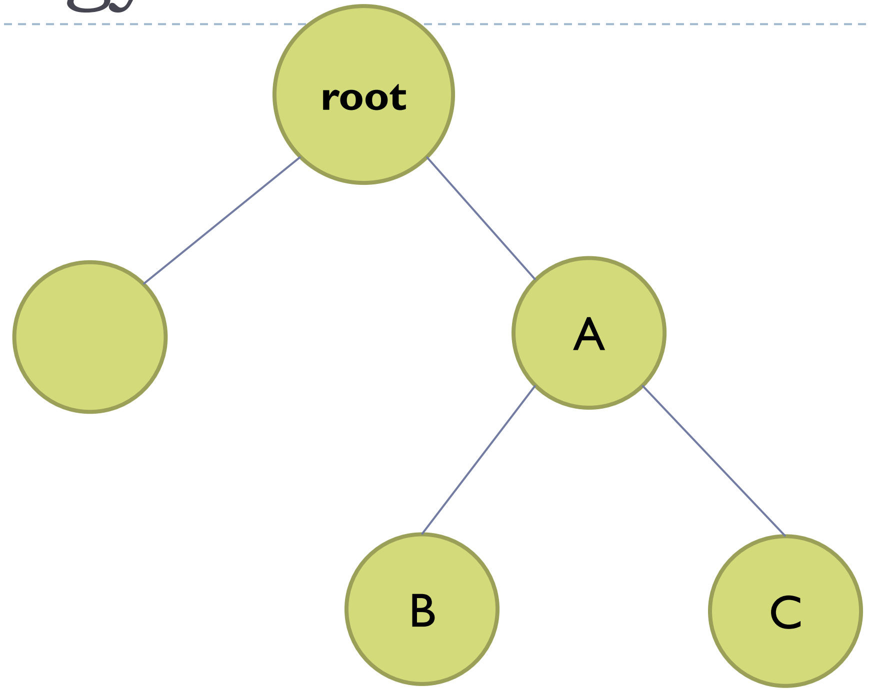


What’s Decision Tree

Predictors				Target
Outlook	**Temp.**	**Humidity**	**Windy**	**Play Golf**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

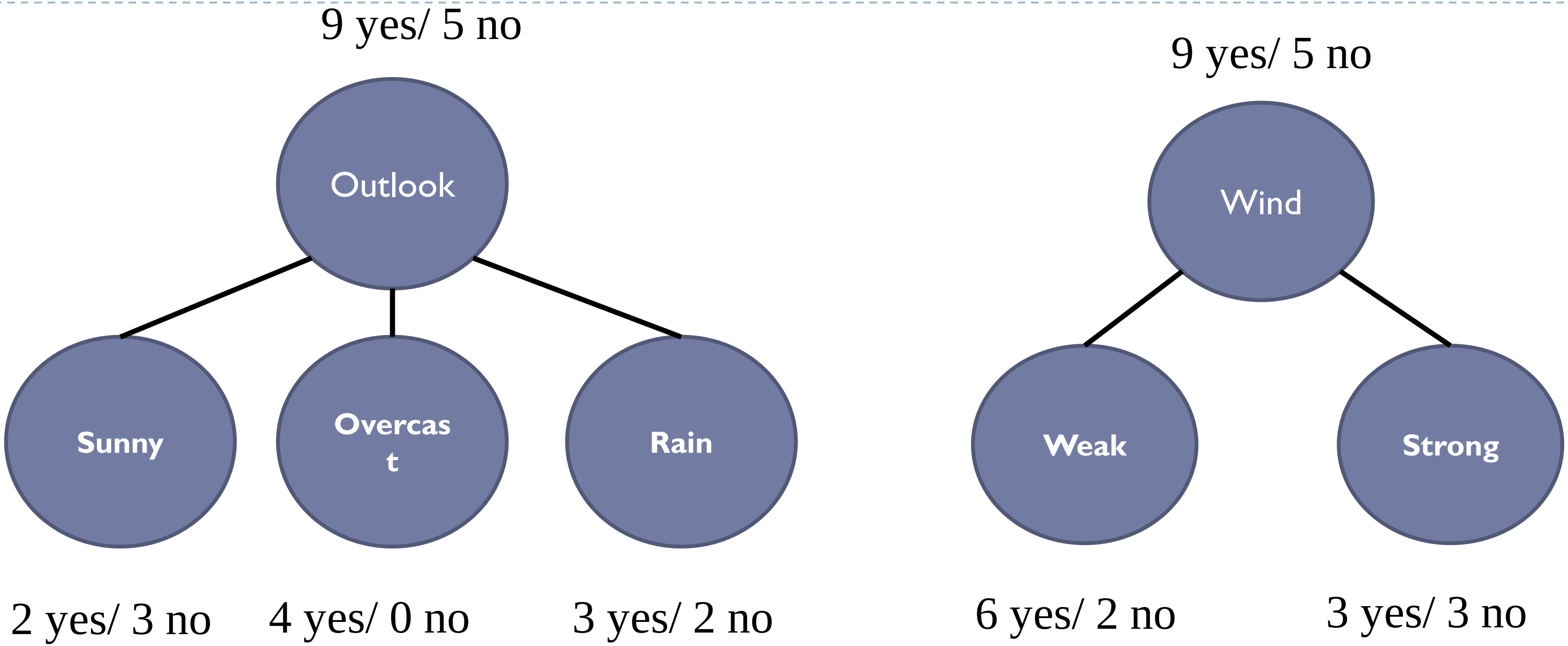


Terminology in Decision Tree



A is parent of B, C
B, C are children of A

How to split on each node?



How to define a good split ?

How to split on each node?

- Information/Gini gain
  - Index to decide how to split each node
  - Usually, we choose max information/gini gain as candidate to split

CART

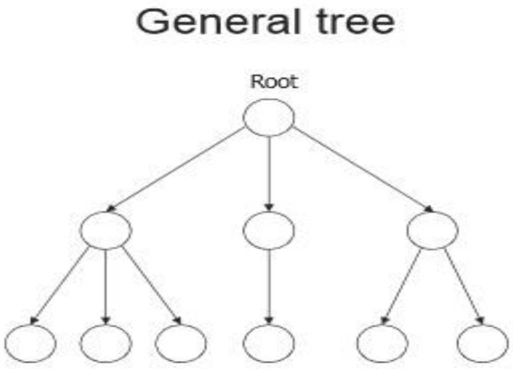
$$Gini\ Gain = Gini(before\ splitting) - E[Gini(after\ splitting)]$$

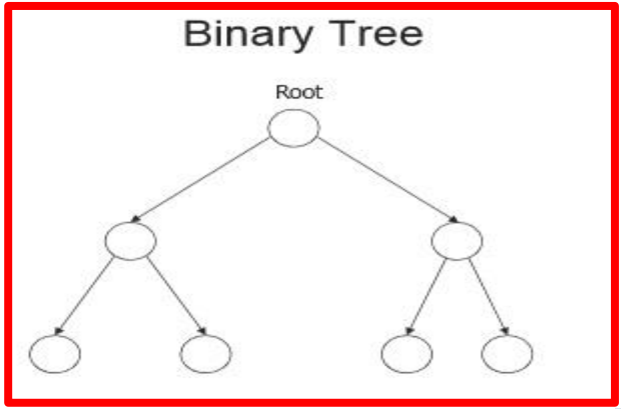
ID3

$$Information\ Gain = Entropy(before\ splitting) - E[Entropy(after\ splitting)]$$

Decision Tree - CART

- Classification and Regression Trees(CART) model is a **binary** tree
- Split Based on One Variable
- Use Gini impurity to define attribute complexity under each feature
- Use Gini gain to split tree





Gini Impurity

J classes and each pi is probability of class i

$$\sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J}(p_i - {p_i}^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} {p_i}^2 = 1 - \sum_{i=1}^{J} {p_i}^2$$

Class 1	0
Class 2	6

$$p(class\ 1) = \frac{0}{6}, \qquad p(class\ 2) = \frac{6}{6}$$

$$Gini = 1 - (\frac{0}{6})^2 - (\frac{6}{6})^2 = 0$$

Class 1	1
Class 2	5

$$p(class\ 1) = \frac{1}{6}, \qquad p(class\ 2) = \frac{5}{6}$$

$$Gini = 1 - (\frac{1}{6})^2 - (\frac{5}{6})^2 = 0.278$$

Class 1	2
Class 2	4

$$p(class\ 1) = \frac{2}{6}, \qquad p(class\ 2) = \frac{4}{6}$$

$$Gini = 1 - (\frac{2}{6})^2 - (\frac{4}{6})^2 = 0.444$$

Example

Gini Large → **Less Purity**

Gini Small → **More Purity**

CART use Gini Gain to Split node

before splitting

Class 1	6
Class 2	6

Gini(before splitting) = 0.5

after splitting

suppose there are two ways (A or B) to split the data

A

Class 1	4
Class 2	3

Gini = 0.489

Class 1	2
Class 2	3

Gini = 0.48

E[Gini(after splitting)]

$$=\frac{7}{12}*0.489+\frac{5}{12}*0.48=0.4852$$

B

Class 1	1
Class 2	4

Gini = 0.32

Class 1	5
Class 2	2

Gini = 0.408

E[Gini(after splitting)]

$$=\frac{5}{12}*0.32+\frac{7}{12}*0.408=0.37$$

CART use Gini Gain to Split node

before splitting

Class 1	**6**
Class 2	**6**

Gini(before splitting) = 0.5

after splitting

Gini Gain on A way
=Gini(before splitting) -E[Gini(after splitting)]
=0.015

Gini Gain on B way
= Gini(before splitting) -E[Gini(after splitting)]
=0.13

Split on B way is better

CART use Gini Gain to Split node

before splitting

Class 1	6
Class 2	6

Gini(before splitting) = 0.5

after splitting

suppose there are two ways (A or B) to split the data

A

Class 1	4
Class 2	3

Class 1	2
Class 2	3

B

Class 1	1
Class 2	4

Class 1	5
Class 2	2

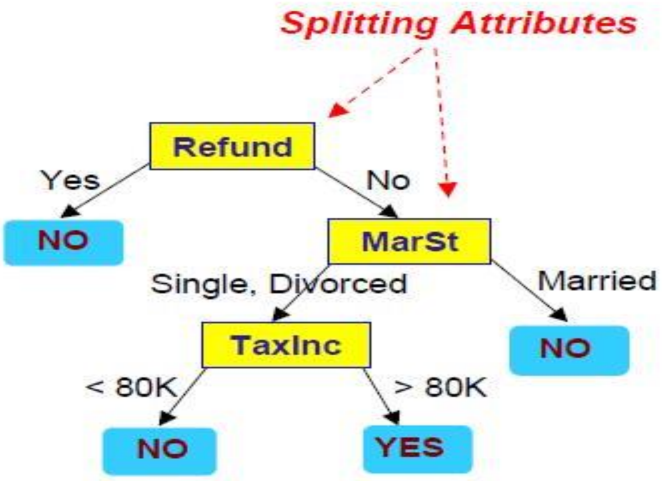
Split on B way is better

Decision Tree – CART Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

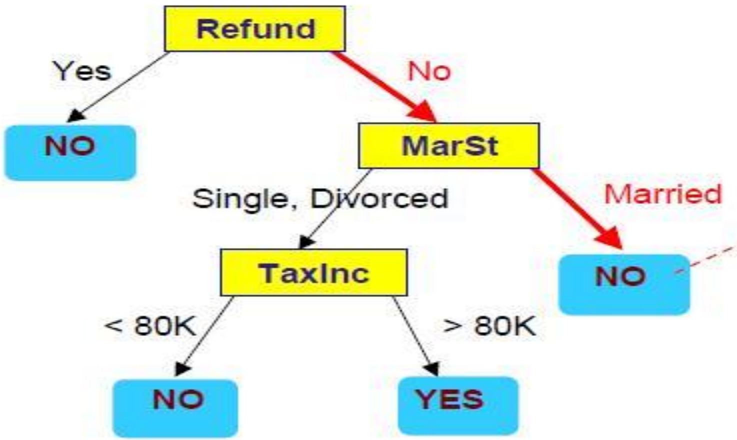
How to deal with continuous attributes

- There are many different way to deal with continuous attributes when building decision tree
  - The most simple way is to split by average of continuous attributes

Decision Tree – CART Example

Test Data

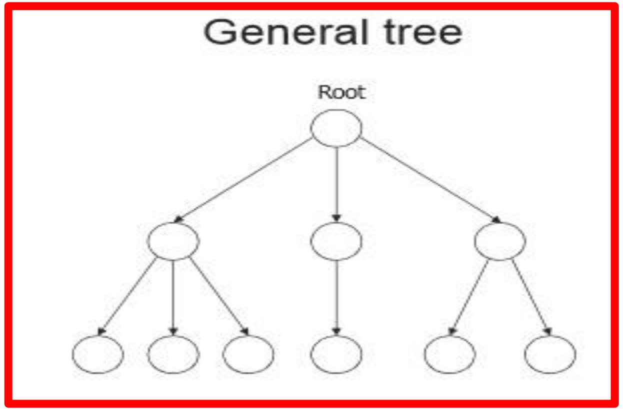
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

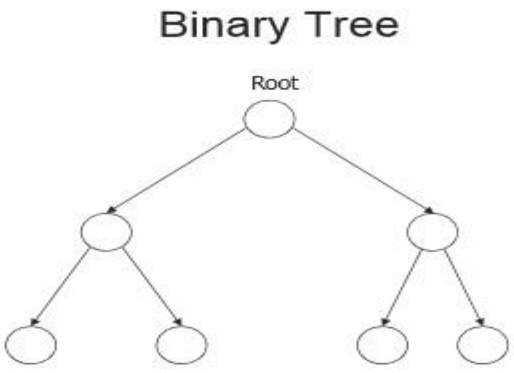


Assign Cheat to “No”

Decision Tree – ID3

- Iterative Dichotomiser 3(ID3) is a famous algorithm to generate decision tree
- Use information gain as index to split each node
- Note that ID3 can split multiple branch at each node





Decision Tree – ID3

- Entropy $E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$

Class 1	0
Class 2	6

$$p(class\ 1) = \frac{0}{6}, \qquad p(class\ 2) = \frac{6}{6}$$

$$Entropy = -0 * \log(0) - 1 * \log(1) = 0$$

Class 1	1
Class 2	5

$$p(class\ 1) = \frac{1}{6}, \qquad p(class\ 2) = \frac{5}{6}$$

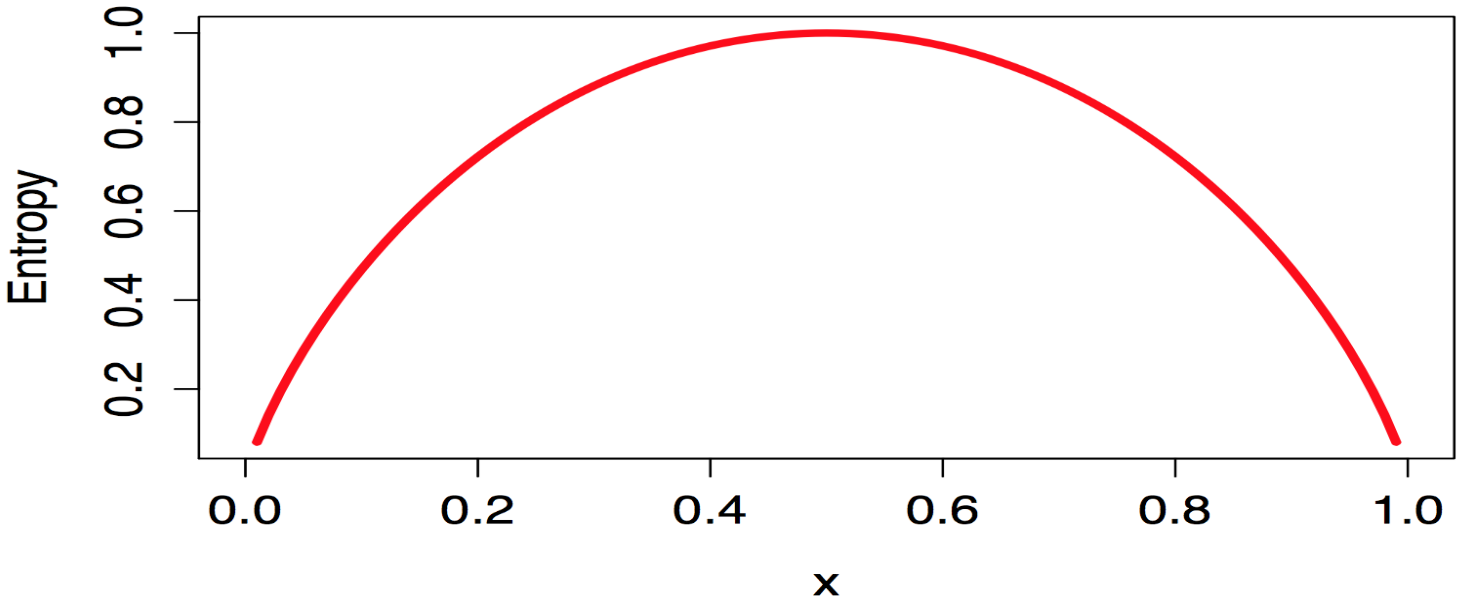
$$Entropy = -\frac{1}{6} * \log\left(\frac{1}{6}\right) - \frac{5}{6} * \log\left(\frac{5}{6}\right) = 0.65$$

Class 1	2
Class 2	4

$$p(class\ 1) = \frac{2}{6}, \qquad p(class\ 2) = \frac{4}{6}$$

$$Entropy = -\frac{2}{6} * \log\left(\frac{2}{6}\right) - \frac{4}{6} * \log\left(\frac{4}{6}\right) = 0.91$$

Decision Tree – ID3



$$Entropy = -x * \log(x) - (1-x) * log(1-x)$$

ID3 use Entropy to Split node

before splitting

Class 1	6
Class 2	6

Entropy(before splitting) = 0.301

after splitting

suppose there are two ways (A or B) to split the data

A

Class 1	4
Class 2	3

Entropy = 0.297

Class 1	2
Class 2	3

Entropy = 0.292

E[Entropy(after splitting)]

$$=\frac{7}{12}*0.297+\frac{5}{12}*0.292=0.294$$

B

Class 1	1
Class 2	4

Entropy = 0.217

Class 1	5
Class 2	2

Entropy = 0.259

E[Entropy(after splitting)]

$$=\frac{5}{12}*0.217+\frac{7}{12}*0.259=0.242$$

ID3 use Entropy to Split node

before splitting

Class 1	**6**
Class 2	**6**

Entropy(before splitting) = 0.5

after splitting

Information Gain on A way
=Entropy(before splitting) -E[Entropy(after splitting)]
=0.007

Information Gain on B way
= Entropy (before splitting) -E[Entropy(after splitting)]
=0.069

Split on B way is better

ID3 use Entropy to Split node

before splitting

Class 1	6
Class 2	6

Gini(before splitting) = 0.5

after splitting

suppose there are two ways (A or B) to split the data

A

Class 1	4
Class 2	3

Class 1	2
Class 2	3

B

Class 1	1
Class 2	4

Class 1	5
Class 2	2

Split on B way is better

Decision Tree – ID3 Example

Predict if playing golf or not

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Decision Tree – ID3 Example

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

$$Entropy(before\ split) = -\frac{5}{14} * \log\left(\frac{5}{14}\right) - \frac{9}{14} * \log\left(\frac{9}{14}\right) = 0.94$$

Decision Tree – ID3 Example

calculate entropy if splitting on **outlook** column

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

$$E[Entropy(after\ splitting)]$$
$$= P(sunny) * E(3,2) + P(overcast) * E(4,0) + P(rainy) * E(2,3)$$
$$=\left(\frac{5}{14}\right) * 0.971 + \left(\frac{4}{14}\right) * 0+\left(\frac{5}{14}\right) * 0.971 = 0.693$$

Decision Tree – ID3 Example

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Information Gain = 0.247			

		Play Golf	
		Yes	No
Temp	Hot	2	2
	Mild	4	2
	Cool	3	1
Information Gain = 0.029			

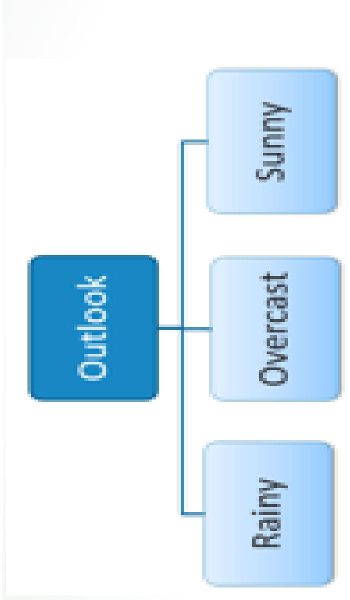
Max Gain

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Information Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Information Gain = 0.048			

Decision Tree – ID3 Example

After first splitting, decision tree look like the following



Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

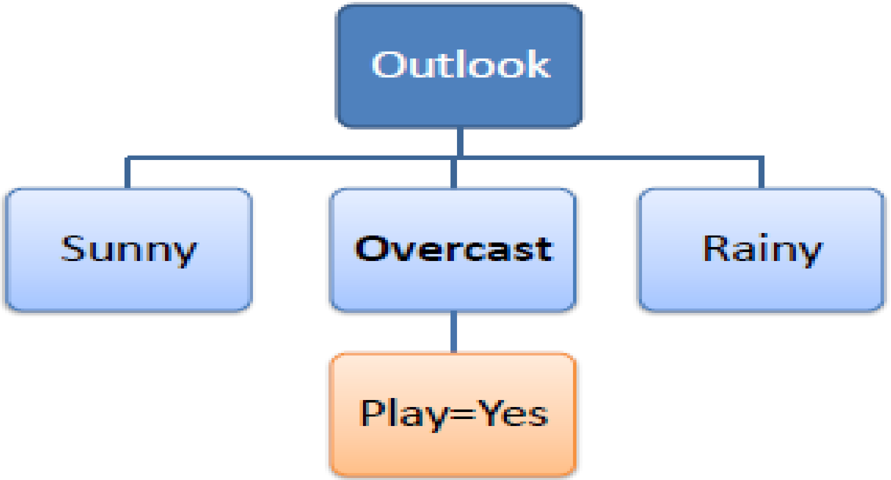
Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes

Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Decision Tree – ID3 Example

No need to further split overcast because all of target are "Yes"

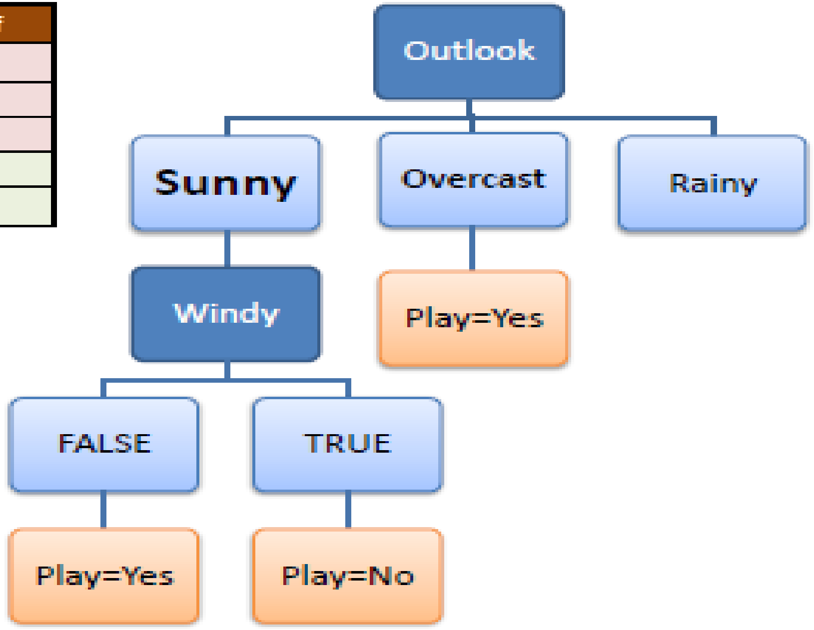
Temp.	Humidity	Windy	Play Golf
Hot	High	FALSE	Yes
Cool	Normal	TRUE	Yes
Mild	High	TRUE	Yes
Hot	Normal	FALSE	Yes



Decision Tree – ID3 Example

Continue split nodes on same method

Temp.	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No



Decision Tree – ID3 Example

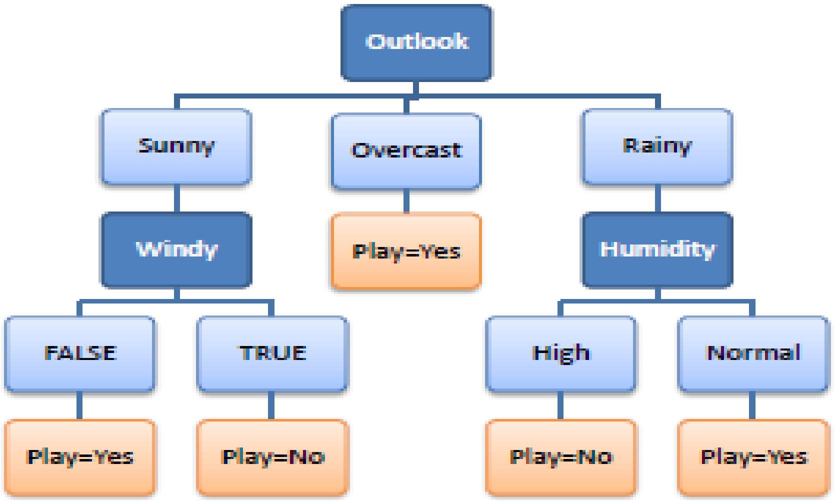
R_1: IF (Outlook=Sunny) AND (Windy=FALSE) THEN Play=Yes

R_2: IF (Outlook=Sunny) AND (Windy=TRUE) THEN Play=No

R_3: IF (Outlook=Overcast) THEN Play=Yes

R_4: IF (Outlook=Rainy) AND (Humidity=High) THEN Play=No

R_5: IF (Outlook=Rain) AND (Humidity=Normal) THEN Play=Yes



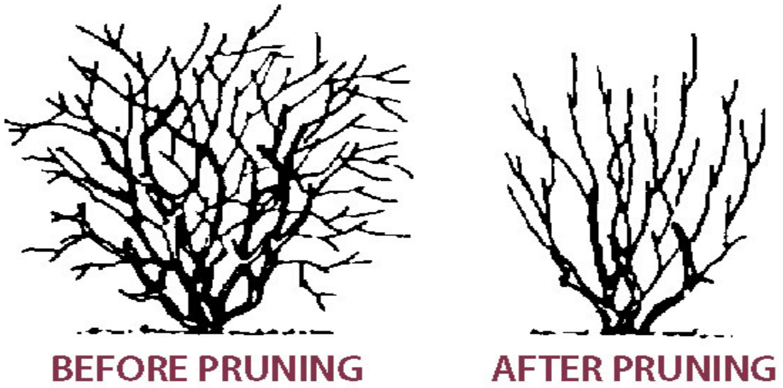
http://www.saedsayad.com/decision_tree.htm

Decision Tree – ID3

- Calculate target Entropy
- Find the information gain on each attribute
- Split tree on an attribute which information gain is max
- Repeat

Pruning

- Pruning is a technique that reduces the size of decision trees
  - Reduce model complexity and overfitting



Stopping Condition

- Pre-pruning
  - Stop the algorithm before it becomes a fully-grown tree
    - Stop if all instances belong to the same class
    - Stop if number of instances is less than some user-specified threshold
    - Stop if expanding the current node does not improve impurity measures
    - ……
- Post-pruning
  - Grow decision tree to its entirety and trim the nodes of the decision tree in a bottom-up
  - If generalization error improves after trimming, replace sub-tree by a leaf node

Example and Practice

- Example
  - Decision Tree (CART)
    - example/supervised learning
- Practice
  - Try to use decision tree to predict if abalone is old or young
    - dataset/abalone.csv
    - practice/supervised learning
    - we assume age > 8 is old and other is young
  - More information about the dataset
    - https://archive.ics.uci.edu/ml/datasets/abalone