Mid-term Exam Stats

- 37 students
  - 100: 3
    - Afzal Hossain
    - Rosaura Vidal Mata
    - Hao Zheng
  - 95-99: 7
  - 90-94: 13
  - 85-89: 10
  - < 85: 4 (min: 75)

- 91 (Mean) $\pm$ 6 (Stdev)
- Median: 92
- Mode: 94

- Get your graded exam paper after class.

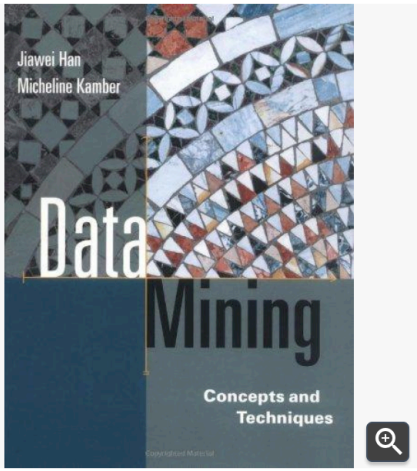
Project Team Pairing

- 12 two-member teams!
- 3 single-member teams:
  - Afzal Hossain (graduate)
  - Rachel Krohn (graduate)
  - Lauren Ferrara (underg.)

❤ Graduates:
- Anselme Mucunguzi
- Satyaki Sikdar
- Famim Talukder
- Kuang Wu
- Shengsheng Yuan

❤ Undergraduates:
- Aron Lam
- Collin Klenke
- David Durkin
- Chris Rho
- Luis Prieb
- Matthew Reily

Best Project Awards (2)



Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)

Jiawei Han; Micheline Kamber

252 ratings by Goodreads

ISBN 10: 1558604898 / ISBN 13: 9781558604896

Published by Morgan Kaufmann, 2000

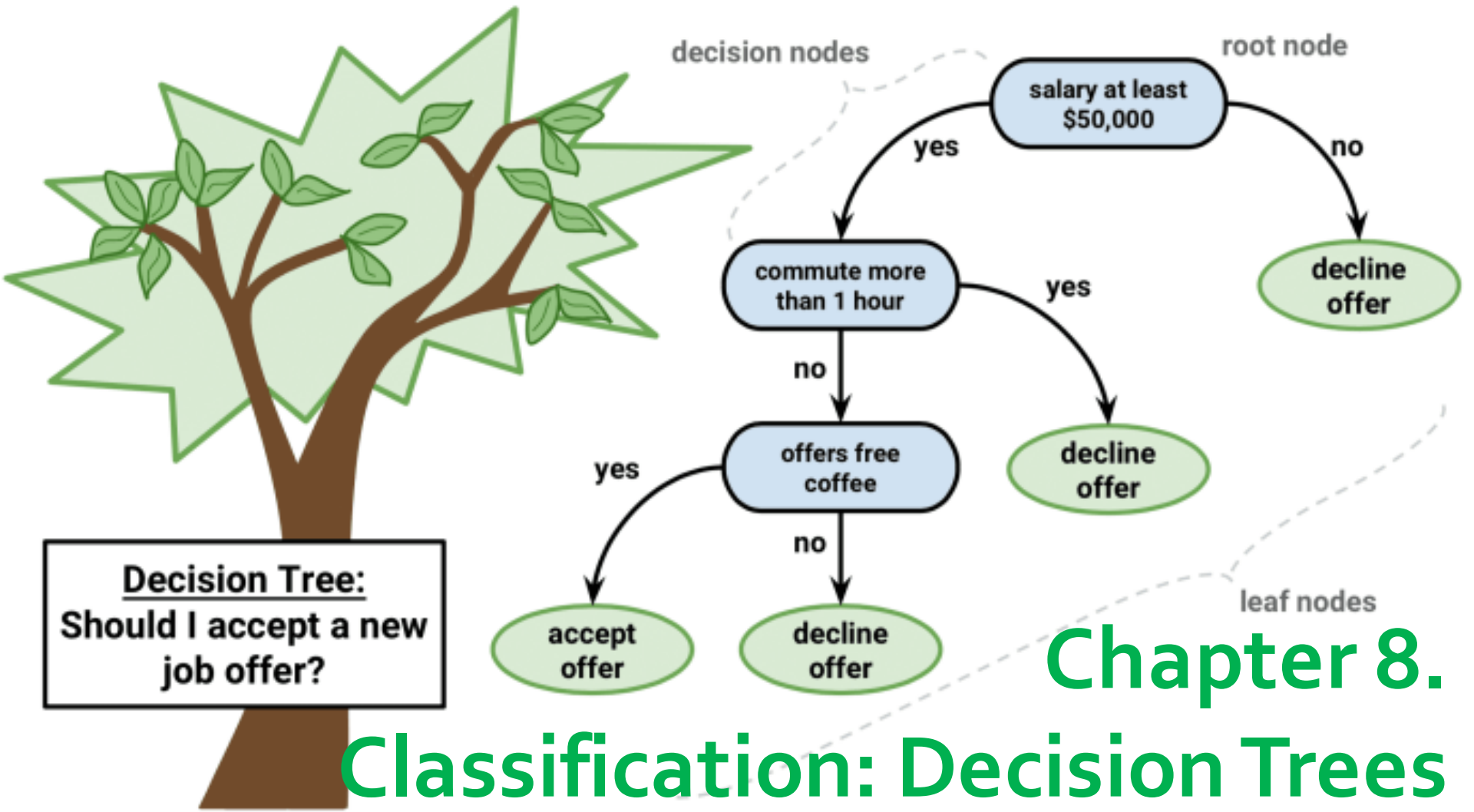
New | Condition: New | Hardcover

2 New from US$ 36.25 | 15 Used from US$ 3.48 | › View all 17 copies of this book

Moving Office Hour?

- Thursday 3:30pm – 4:30pm

- Other options:
  - Wednesday 10am – 11am
  - Wednesday 11am – 12pm
  - Thursday 10am – 11am



Chapter 8. Classification: Decision Trees

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

Supervised vs. Unsupervised Learning

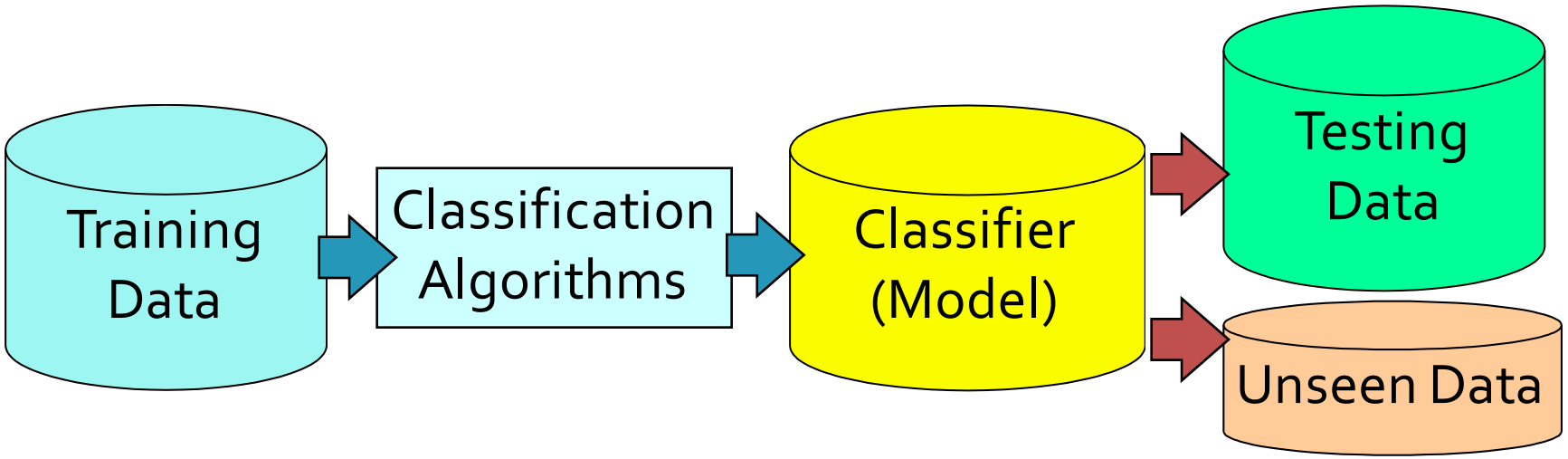
- Supervised learning (**classification**)
  - Supervision: The *training* data instances and their attributes/features are accompanied by labels indicating the class of the instances.
  - Predict class labels for *testing* data instances.
- Unsupervised learning (**clustering**)
  - The class labels of training data is unknown
  - Given a set of attributes, with the aim of establishing the existence of classes or clusters.

Classification: Applications

- Credit/loan approval
- Medical diagnosis: if a tumor is cancerous or benign
- Fraud detection: if a transaction is fraudulent
- Web page categorization: which category it is
- …

Classification: A Two-Step Process

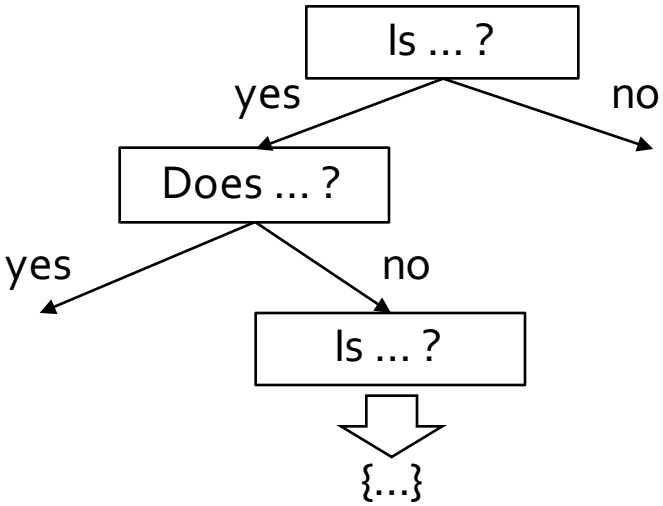
- **Model construction**
  - **Models: Decision trees, Naïve Bayes, SVM, Neural Networks, etc.**
- **Model usage**
  - Estimate accuracy of the model
    - Accuracy: % of test instances that are correctly classified
    - Test set is independent of training set (otherwise **overfitting**)
  - If the accuracy is acceptable, use the model to classify new data



How to Construct a Decision-Trees Model for Classification?

- Let's play a game first!

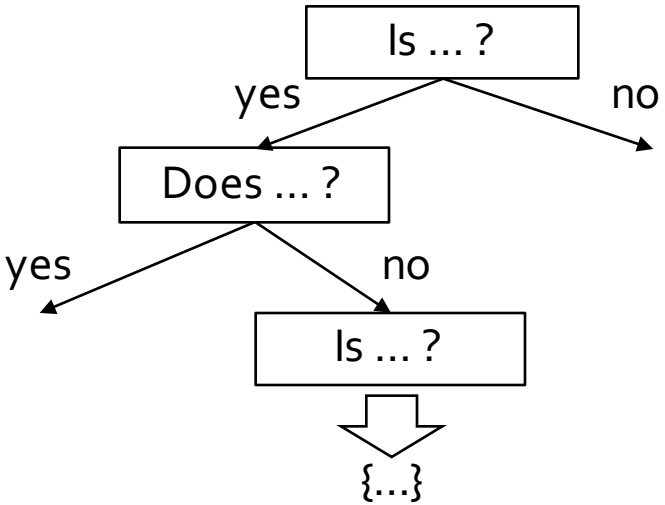
{Barack Obama, Hillary Clinton, Ellen DeGeneres, Abraham Lincoln, Superman, …}



Decision Trees

- A directed tree structure comprised of nodes
- Each **node** specified an evaluation on a feature
- Each **branch** corresponds to a feature value
- Each **leaf** signified a categorical decision or class

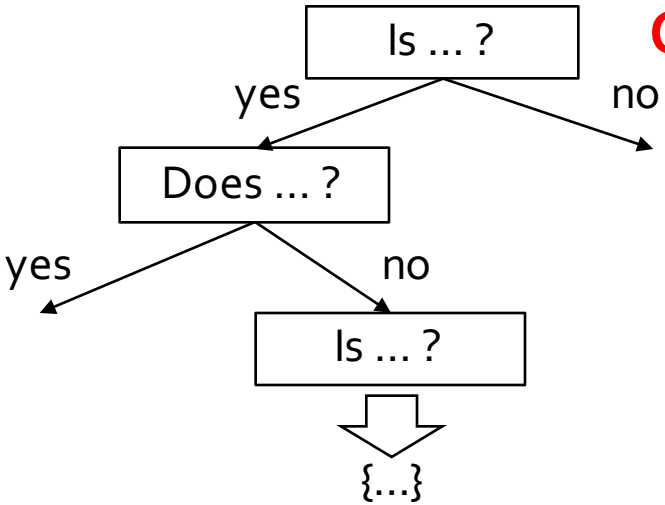
{Barack Obama, Hillary Clinton, Ellen DeGeneres, Abraham Lincoln, Superman, …}



Constructing Decision Trees

- Top down, recursive divide-and-conquer
- Select best feature for root node
- Construct branch for each possible feature value
- Split data into mutually exclusive subsets along each branch
- **Repeat** procedure recursively for each branch
- Terminate into leaf node after adequate performance

{Barack Obama, Hillary Clinton, Ellen DeGeneres, Abraham Lincoln, Superman, …}



Q: Which feature to select?

PRINCIPLE

- **Reduce uncertainty** as much as possible

max ReducedUncertainty(Y|X)

= Uncertainty({instances at the root node})

– Uncertainty({instances at child nodes*|selected_attribute})

*child nodes: values of the selected attributes

Revisiting Entropy

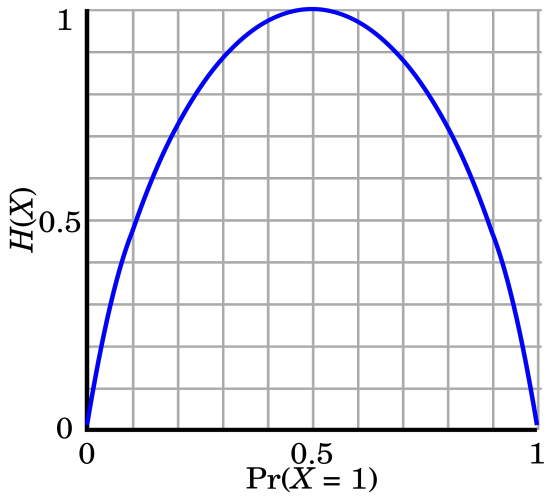
- Entropy (Information Theory)
  - A measure of **uncertainty** associated with a random number
  - Calculation: For a discrete random variable Y taking m distinct values $\{y_1, y_2, \ldots, y_m\}$

$$H(Y) = -\sum_{i=1}^{m} p_i \log(p_i) \quad where\ p_i = P(Y = y_i)$$

$H(Pr(X=1)=0.5) = -0.5\log_2(0.5) - 0.5\log_2(0.5) = 1$

$H(Pr(X=1)=0 \text{ or } 1) = -0\log_2(\varepsilon) - 1\log_2(1-\varepsilon) = 0$ $(\varepsilon \rightarrow 0)$

  - Interpretation
    - Higher entropy → higher uncertainty
    - Lower entropy → lower uncertainty



m = 2

- Conditional entropy

$$H(Y|X) = \sum_{x} p(x) H(Y|X = x)$$

Information Gain

- Can be defined as expected reduction in entropy by partitioning set of instances according to feature X:

$$max\_X \quad IG(Y|X) = H(Y) - H(Y|X)$$

Information gain of class Y given feature X

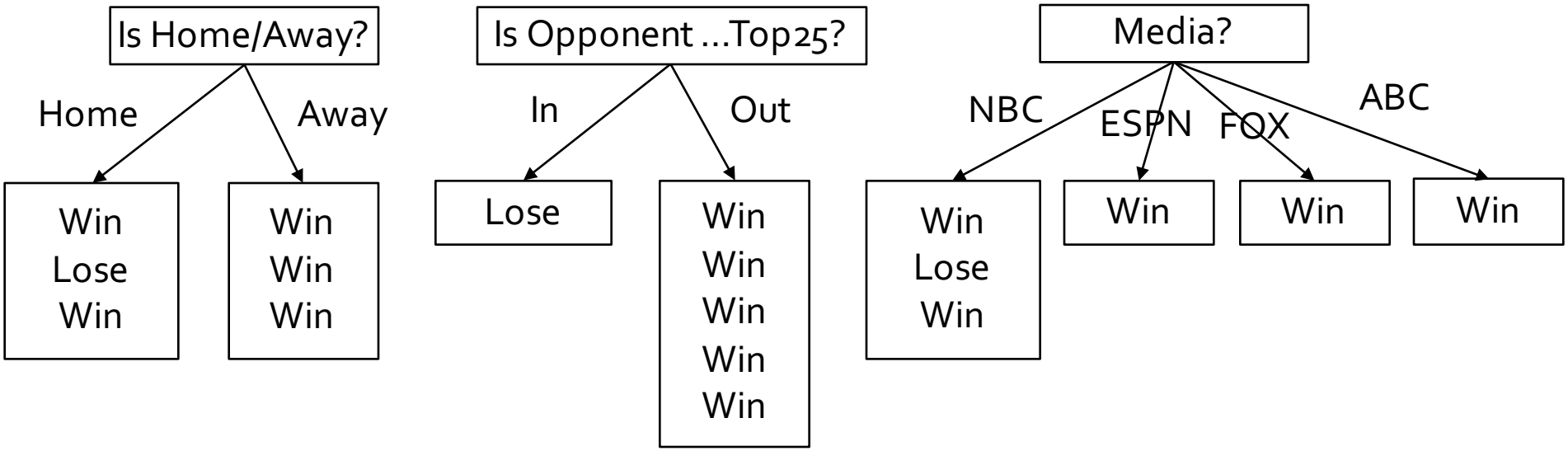
Unconditional entropy of class Y

Conditional entropy of class Y given feature X

Game Classification (Result Prediction)

			Is Home/Away?	Is Opponent in AP Top 25 at Preseason?	Media	Label: Win/Lose
1	9/2/17	Temple	Home	Out	1-NBC	Win
2	9/9/17	Georgia	Home	In	1-NBC	Lose
3	9/16/17	Boston College	Away	Out	2-ESPN	Win
4	9/23/17	Michigan State	Away	Out	3-FOX	Win
5	9/30/17	Miami Ohio	Home	Out	1-NBC	Win
6	10/7/17	North Carolina	Away	Out	4-ABC	Win

Partitioning set of instances according to feature



Information Gain Calculation

Y = {Win*5, Lose*1}

H(Y) = $-\frac{5}{6}\log_2\frac{5}{6}-\frac{1}{6}\log_2\frac{1}{6}=0.65$

X1_HomeAway = {Home*3, Away*3}

H(Y|X1) = H(Y|Home)+H(Y|Away)

$=\frac{3}{6}\times\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)+\frac{3}{6}\times(-\frac{3}{3}\log_2\frac{3}{3})$

= 0.5×0.92+0 = 0.46

IG(Y|X1) = 0.65-0.46 = 0.19

Information Gain Calculation

Y = {Win*5, Lose*1}

H(Y) = $-\frac{5}{6}\log_2\frac{5}{6}-\frac{1}{6}\log_2\frac{1}{6}=0.65$

X2_Top25 = {In*1, Out*5}

H(Y|X2) = H(Y|In)+H(Y|Out)

$$=\frac{1}{6}\times\left(-\frac{1}{1}\log_2\frac{1}{1}\right)+\frac{5}{6}\times\left(-\frac{5}{5}\log_2\frac{5}{5}\right)$$

= 0

IG(Y|X2) = 0.65-0 = 0.65

Information Gain Calculation

Y = {Win*5, Lose*1}

H(Y) = $-\frac{5}{6}\log_2\frac{5}{6}-\frac{1}{6}\log_2\frac{1}{6}=0.65$

X3_Media = {NBC*3, ESPN*1, FOX*1, ABC*1}

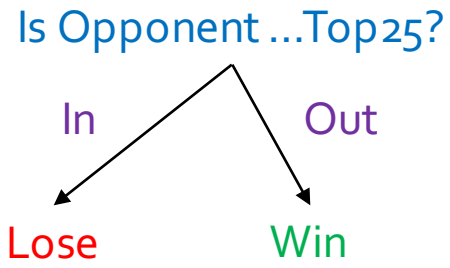
H(Y|X3) = H(Y|NBC)+H(Y|ESPN)+H(Y|FOX)+H(Y|ABC)

$=\frac{3}{6}\times\left(-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}\right)+\frac{1}{6}\times(-\frac{1}{1}\log_2\frac{1}{1})+\frac{1}{6}\times(-\frac{1}{1}\log_2\frac{1}{1})+\frac{1}{6}\times(-\frac{1}{1}\log_2\frac{1}{1})$

= 0.5*0.92+0+0+0 = 0.46

IG(Y|X3) = 0.65-0.46 = 0.19

Final Decision Tree



Test:

1	10/21/17	USC	Home	In	1-NBC	Lose
2	10/28/17	North Carolina State	Home	Out	1-NBC	Win
3	11/4/17	Wake Forest	Home	Out	1-NBC	Win
4	11/18/17	Navy	Home	Out	1-NBC	Win

I don't think this is a good classifier…

(overfitting)

I bet 4 straight wins so the accuracy is 75%!

Quinlan's Example (1986): Playing Tennis

	Outlook	Temperature	Humidity	Windy	Label: Play?
1	Sunny	Hot	High	"False"	No
2	Sunny	Hot	High	"True"	No
3	Overcast	Hot	High	"False"	Yes
4	Rainy	Mild	High	"False"	Yes
5	Rainy	Cool	Normal	"False"	Yes
6	Rainy	Cool	Normal	"True"	No
7	Overcast	Cool	Normal	"True"	Yes
8	Sunny	Mild	High	"False"	No
9	Sunny	Cool	Normal	"False"	Yes
10	Rainy	Mild	Normal	"False"	Yes
11	Sunny	Mild	Normal	"True"	Yes
12	Overcast	Mild	High	"True"	Yes
13	Overcast	Hot	Normal	"False"	Yes
14	Rainy	Mild	High	"True"	No
1	Rainy	Hot	High	"False"	?

Information Gain Calculation

Y = {Yes*9, No*5}

H(Y) = $-\frac{9}{14}\log_2\frac{9}{14}-\frac{5}{14}\log_2\frac{5}{14}=0.94$

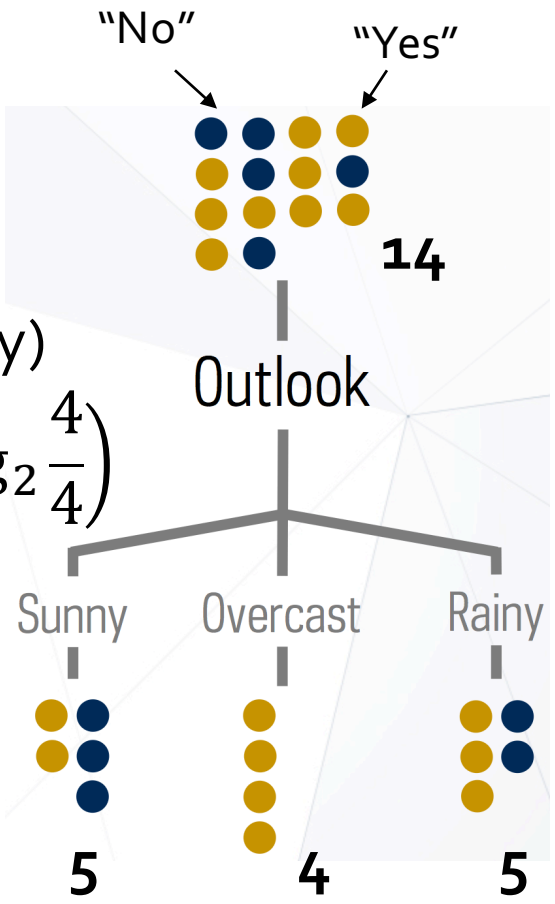
X1_Outlook = {Sunny*5, Overcast*4, Rainy*5}

H(Y|X1) = H(Y|Sunny)+H(Y|Overcast)+H(Y|Rainy)

$$=\frac{5}{14}\times\left(-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}\right)+\frac{4}{14}\times\left(-\frac{4}{4}\log_2\frac{4}{4}\right)+\frac{5}{14}\times\left(-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}\right)$$

= 0.345+0+0.345 = 0.69

IG(Y|X1) = 0.94-0.69 = 0.25



Information Gain Calculation

IG(Y|X1_Outlook) = 0.25

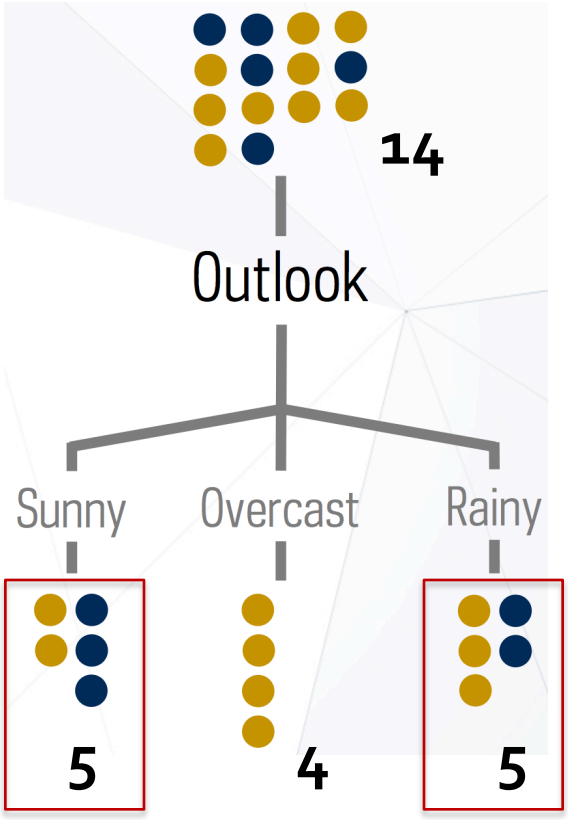
IG(Y|X2_Temperature) = 0.03

IG(Y|X3_Humidity) = 0.15

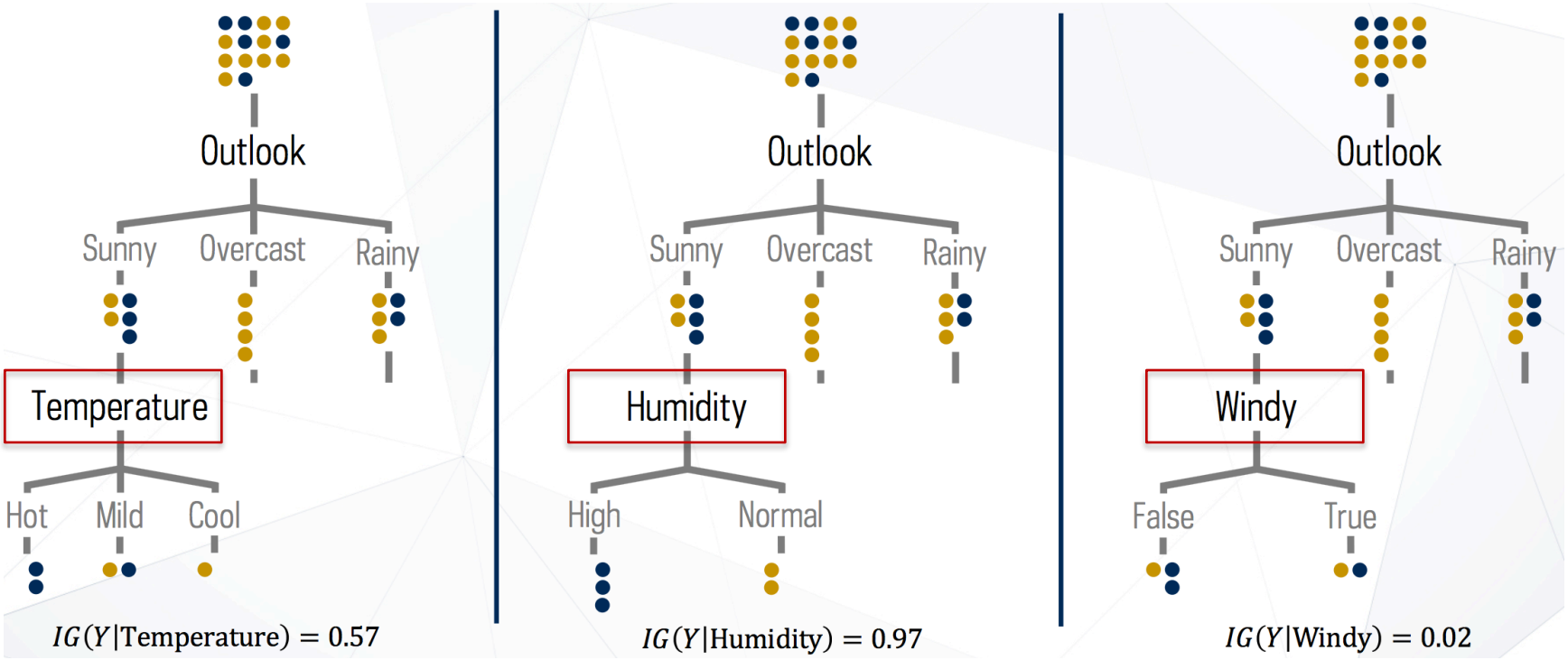
IG(Y|X4_Windy) = 0.05

So the best feature is Outlook.

What's next step?

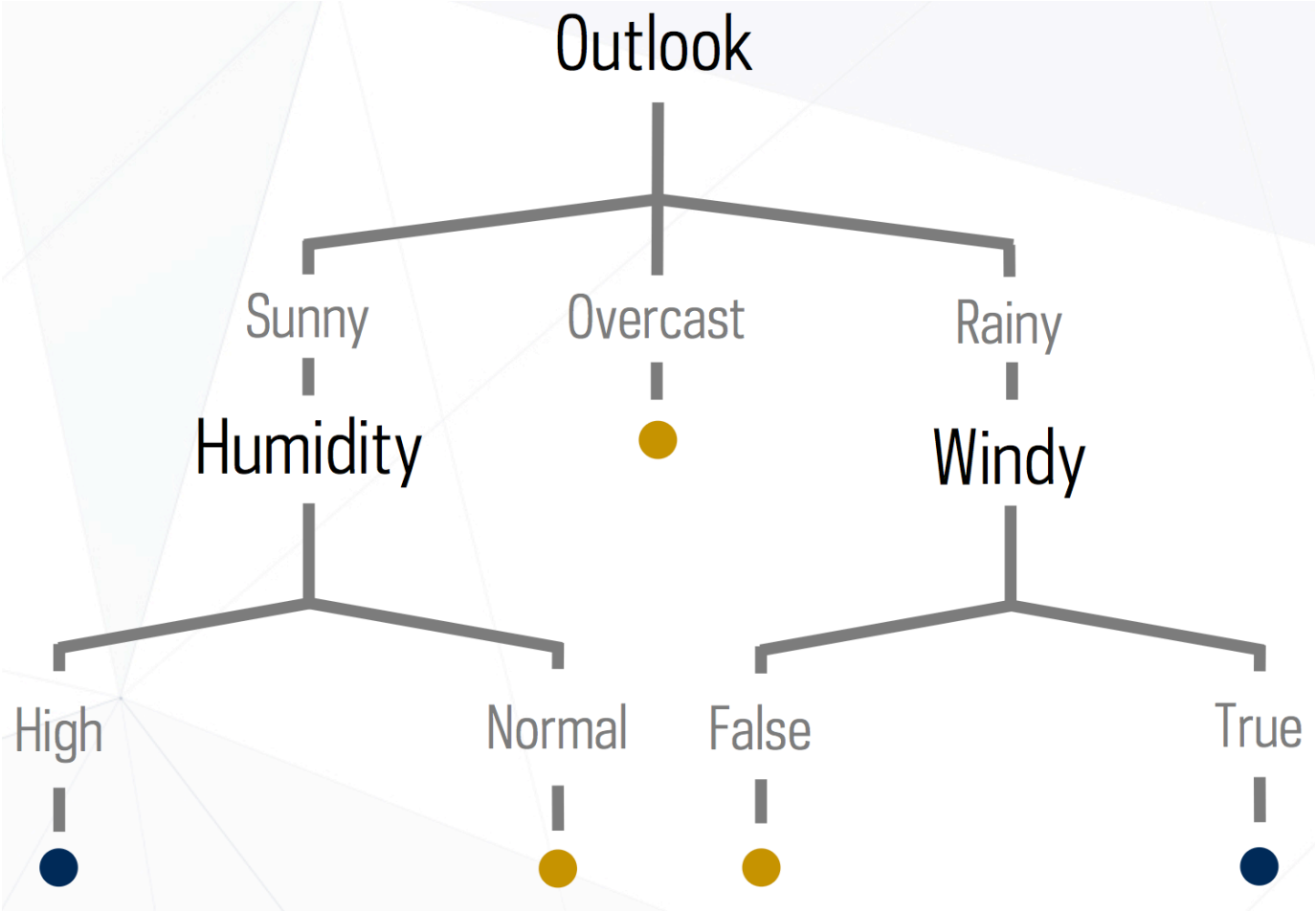


Next Step



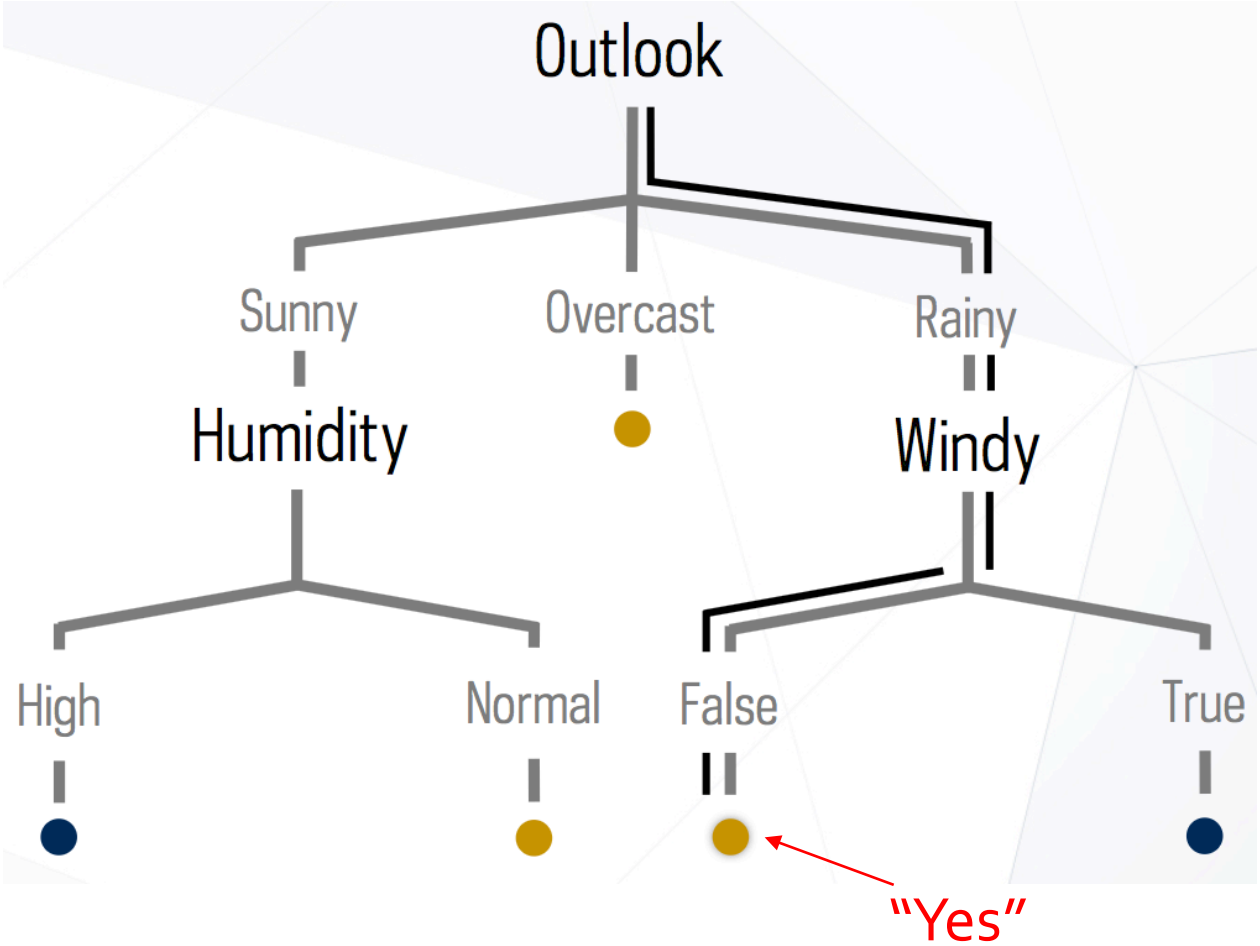
Good!

Final Decision Tree



Prediction

1	Rainy	Hot	High	"False"	?

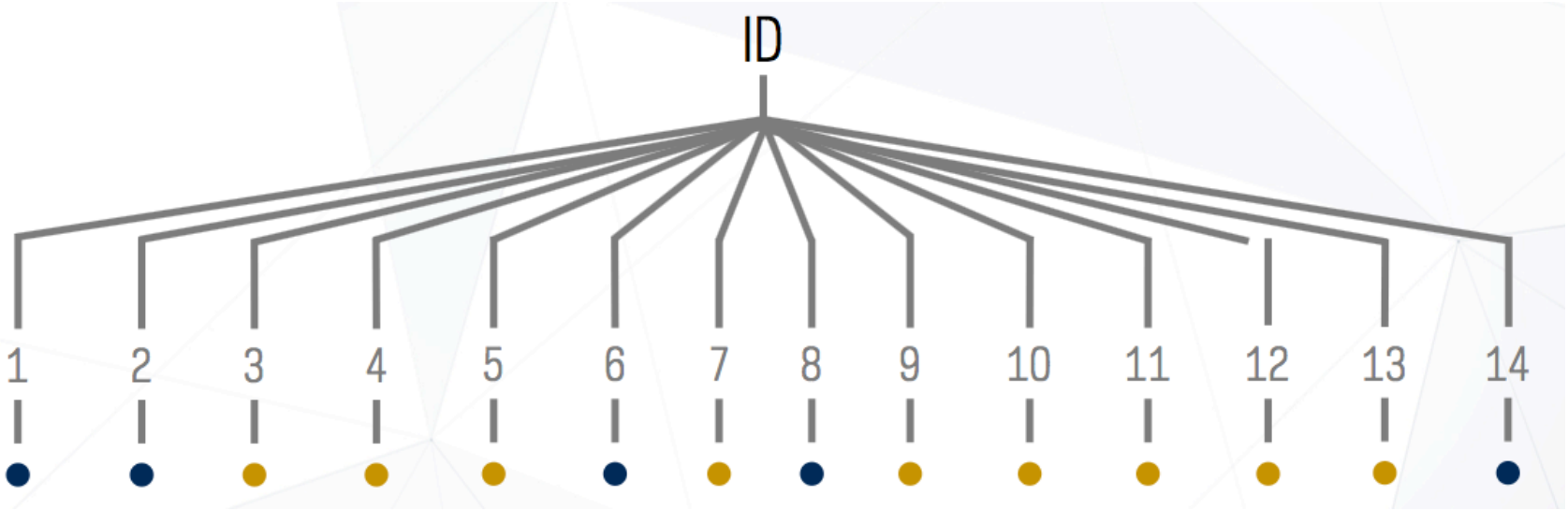


Splitting Criterion

- Information Gain (used in ID3)
  - Iterative Dichotomiser 3 invented by Ross Quinlan in 1986
- Gain Ratio (used in C4.5)
  - C4.5 is an extension of Quinlan's earlier ID3 algorithm, developed by Ross Quinlan
  - It became quite popular after ranking #1 in the Top 10 Algorithms in Data Mining pre-eminent paper published by Springer LNCS in 2008
- Gini Measure (used in CART)
  - Classification And Regression Trees by Breiman et al. in 1984

Gain Ratio (C4.5)

- Information gain measure is biased towards **highly-branching attributes** = with a large number of values
- Entropy of splitting on "ID" is 0. IG for "ID" is maximal.



Gain Ratio (C4.5)

- Corrects information by calculating the *intrinsic information* of a split
  - Information needed to identify branch
  - Accounts for number and size of branches
- Given entropy of instances distributed into branches

$$SplitInfo(S,F) = -\sum \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

- Gain ratio is defined as

$$GainRatio(S,F) = \frac{IG(S,F)}{SplitInfo(S,F)}$$

S: "samples"
F: feature

Gain Ratio Calculation

IG(Y|X1_Outlook) = 0.25

IG(Y|X2_Temperature) = 0.03

IG(Y|X3_Humidity) = 0.15

IG(Y|X4_Windy) = 0.05

SplitInfo(X2_Temperature)

$$= -\frac{4}{14}\log_2\frac{4}{14} - \frac{6}{14}\log_2\frac{6}{14} - \frac{4}{14}\log_2\frac{4}{14} = 1.56$$

GainRatio(X2_Temperature)

=0.03/1.56 = **0.02**

Temperature
Hot
Hot
Hot
Mild
Cool
Cool
Cool
Mild
Cool
Mild
Mild
Mild
Hot
Mild

Gini Index (CART)

- Another splitting criteria. Defined as

$$Gini = 1 - \sum_{i=1}^{K} p_k^2$$

where p_k denotes the proportion of instances belonging to class k (k = 1… K).

Compared with Information Entropy (Info, or H):

$$Info = H = -\sum_{i=1}^{K} p_k \log p_k$$

IG vs Gini

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

$$Gain(A) = Info(D) - Info_A(D)$$

$$\Delta gini(A) = gini(D) - gini_A(D)$$

Maximize

Gini Index Calculation

Y = {Yes*9, No * 5}

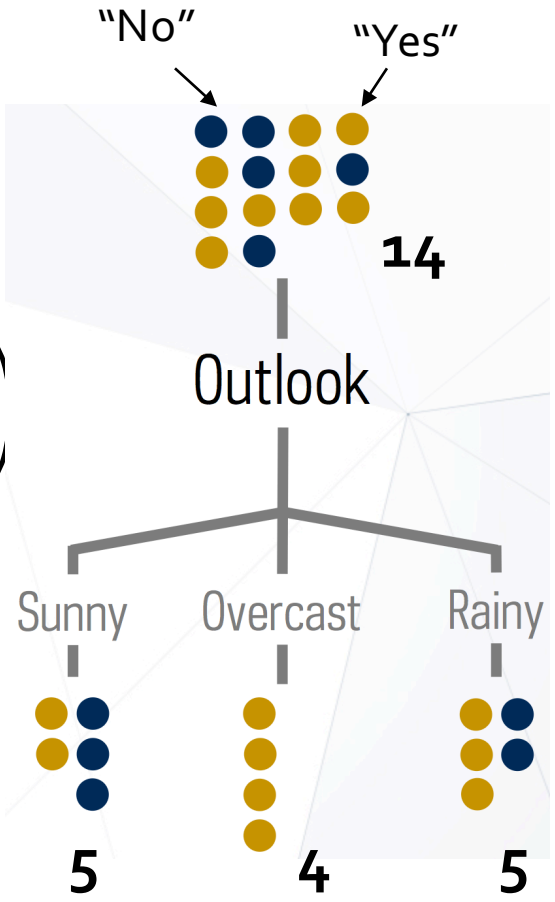
$$\text{Gini}(Y) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.46$$

X1_Outlook = {Sunny*5, Overcast*4, Rainy*5}

$$\text{Gini}(Y|X1) = \frac{5}{14} \times \left(1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right) + \frac{4}{14} \times \left(1 - \left(\frac{4}{4}\right)^2\right)$$

$$+ \frac{5}{14} \times \left(1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right) = 0.34$$

$$\Delta\text{Gini}(Y|X1) = 0.46 - 0.34 = 0.12$$



Gini Index Calculation

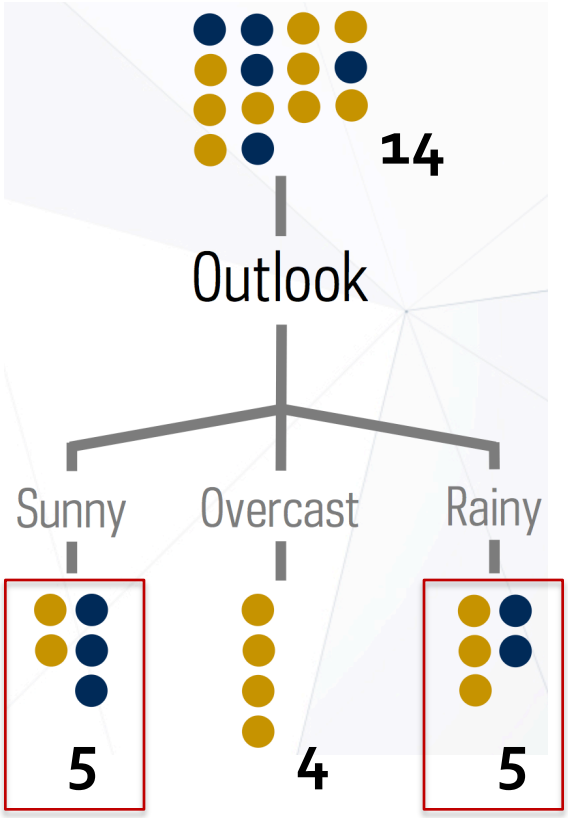
Δgini (Y|X1_Outlook) = 0.12

Δgini(Y|X2_Temperature) = 0.02

Δgini(Y|X3_Humidity) = 0.09

Δgini(Y|X4_Windy) = 0.03

So the best feature is Outlook.



Decision Tree Demo

http://www.meng-jiang.com/teaching/DecisionTreeDemo.zip

This is for Question 1 in your exercise ☺

References

- C. Apte and S. Weiss. Data mining with decision trees and decision rules. Future Generation Computer Systems, 13, 1997
- P. K. Chan and S. J. Stolfo. Learning arbiter and combiner trees from partitioned data for scaling machine learning. KDD'95
- A. J. Dobson. An Introduction to Generalized Linear Models. Chapman & Hall, 1990.
- R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification, 2ed. John Wiley, 2001
- U. M. Fayyad. Branching on attribute values in decision tree generation. AAAI'94.
- Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. J. Computer and System Sciences, 1997.
- J. Gehrke, R. Ramakrishnan, and V. Ganti. Rainforest: A framework for fast decision tree construction of large datasets. VLDB'98.
- J. Gehrke, V. Gant, R. Ramakrishnan, and W.-Y. Loh, BOAT -- Optimistic Decision Tree Construction. SIGMOD'99.
- T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, 2001.
- T.-S. Lim, W.-Y. Loh, and Y.-S. Shih. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning, 2000

References (cont.)

- J. Magidson. The Chaid approach to segmentation modeling: Chi-squared automatic interaction detection. In R. P. Bagozzi, editor, Advanced Methods of Marketing Research, Blackwell Business, 1994
- M. Mehta, R. Agrawal, and J. Rissanen. SLIQ : A fast scalable classifier for data mining. EDBT'96
- T. M. Mitchell. Machine Learning. McGraw Hill, 1997
- S. K. Murthy, Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery 2(4): 345-389, 1998
- J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81-106, 1986.
- J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
- J. R. Quinlan. Bagging, boosting, and c4.5. AAAI'96.
- R. Rastogi and K. Shim. Public: A decision tree classifier that integrates building and pruning. VLDB'98
- J. Shafer, R. Agrawal, and M. Mehta. SPRINT : A scalable parallel classifier for data mining. VLDB'96
- J. W. Shavlik and T. G. Dietterich. Readings in Machine Learning. Morgan Kaufmann, 1990
- P. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining. Addison Wesley, 2005
- S. M. Weiss and C. A. Kulikowski. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufman, 1991
- S. M. Weiss and N. Indurkhya. Predictive Data Mining. Morgan Kaufmann, 1997
- I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques, 2ed. Morgan Kaufmann, 2005