Lecture Notes on Pattern Recognition

Session 01(B): Bayesian ML Methods

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

0

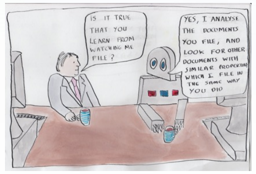
Road map

- Data Analytics
  - Machine Learning
    - Overview
  - Supervised Learning
    - Naive Bayes Classifier
      - Training
      - Prediction
    - Decision Tree
      - Entropy
      - Information Gain

1

Machine Learning

- Supervised Learning
- Unsupervised Learning



2

Intelligent Applications

Did a Human or a Computer Write This?

A shocking amount of what we're reading is created not by humans, but by computer algorithms. Can you tell the difference? Take the quiz.

MARCH 7, 2015

1 "A shallow magnitude 4.7 earthquake was reported Monday morning five miles from Westwood, California, according to the U.S. Geological Survey. The temblor occurred at 6:25 a.m. Pacific time at a depth of 5.0 miles."

Human
Computer

2 "Apple's holiday earnings for 2014 were record shattering. The company earned an $18 billion profit on $74.6 billion in revenue. That profit was more than any company had ever earned in history."

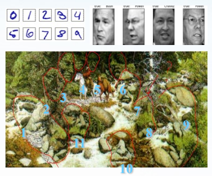
Human
Computer

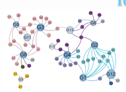
The New York Times

3

Intelligent Applications

- What digit is it?
  - zip code in US Post
- Where are the faces?
  - face detection
- Whose face is it?
  - face recognition
- Where are the groups?
  - Community detection





4

Intelligent Applications

Watson (computer)

From Wikipedia, the free encyclopedia

"IBM Watson" redirects here. For the laboratory, see Thomas J. Watson

Watson is an artificially intelligent computer system capable of answering questions posed in natural language,[2] developed in IBM's DeepQA project by a research team led by principal investigator David Ferrucci. Watson was named after IBM's Thomas J. Watson.[3][4] The computer system was specifically developed to answer questions on the quiz show *Jeopardy!*.[5] In 2011, Watson competed on *Jeopardy!* against former winners Brad Rutter and Ken Jennings.[3][6] Watson received the first prize of $1 million.[7]

Watson, Ken Jennings, and Brad Rutter in their *Jeopardy!* exhibition match.

Watson's avatar, inspired by the IBM "smarter planet" logo[1]

Stanley (vehicle)

From Wikipedia, the free encyclopedia

Stanley is an autonomous car created by Stanford University's Stanford Racing Team in cooperation with the Volkswagen Electronics Research Laboratory (ERL). It competed in, and won, the 2005 DARPA Grand Challenge,[1] earning the Stanford Racing Team the 2 million dollar prize.

Stanley parked after the 2005 DARPA Grand Challenge

5

Intelligent Applications

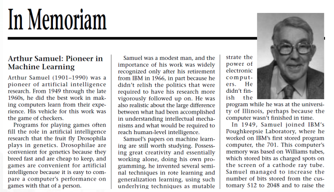
- **AlphaGo** is a computer program developed by Google DeepMind in London to play the board game Go.[1] AlphaGo
  - In March 2016, it beat Lee Sedol in a five-game match, the first time a computer Go program has beaten a professional without handicaps

6

Machine Learning

- ***"Field of study that gives computers the ability to learn without being explicitly programmed"***

Arthur Samuel



7

Machine Learning

- ***"Field of study that gives computers the ability to learn without being explicitly programmed"***

Arthur Samuel

- *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"*

Tom Mitchell

8

Machine Learning

- *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"*

Tom Mitchell

Watson, Ken Jennings, and Brad Rutter in their Jeopardy! exhibition match.

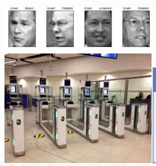
Experience E	Task T	Performance P
databases of millions of question-answer pairs	given an question, find the best answer	how accurate the answer is

9

Machine Learning

- *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"*

Tom Mitchell



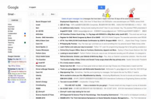
Experience E	Task T	Performance P
databases of thousands of known faces	given a new photo, recognise the name of the face	how accurate the recognition is

10

Machine Learning

- *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"*

Tom Mitchell



Experience E	Task T	Performance P
Examples of spam emails and not-spam email	To assign a label "spam" or "not-spam" to an email	how accurate spam email can be detected

11

Machine Learning

- *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"*

Tom Mitchell

Experience E	Task T	Performance P
Credit-card transactions deemed as fraud and not-fraud	To assign 'fraud' or "not fraud" to a given credit-card transaction	how accurate a credit-card fraud transaction can be detected.

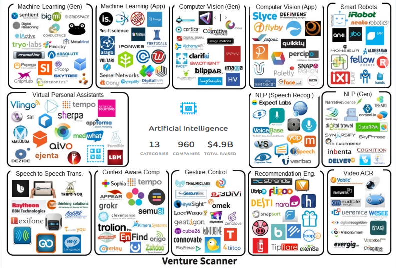
12

Why Machine Learning in Data Science?

- Growing flood of data – big data, which is impossible to handle by human.
  - 1 trillion webpages per day
  - 187 billions emails exchanged per day, etc….
- Computational power is available easier than ever.
  - Lots of progress in algorithms and theory.
  - Solving problems of societal impacts, humanity and nature.
  - "Budding" industry, start-ups

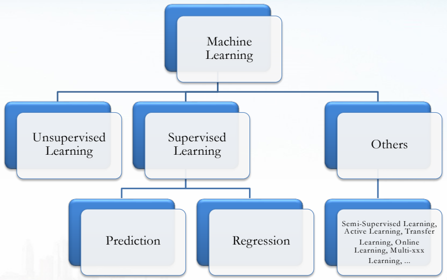
13

Why Machine Learning in Data Science?



14

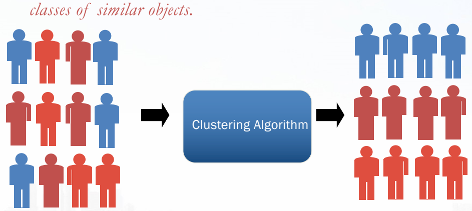
Supervised vs Unsupervised Machine Learning



15

Supervised vs Unsupervised Machine Learning

- Unsupervised Learning, aka **Clustering**
  - is *the process of grouping a set of physical or abstract objects into classes of similar objects.*



16

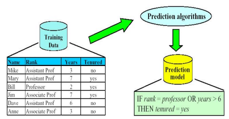
Supervised vs Unsupervised Machine Learning

- Supervised Learning, aka **Prediction (Predictive Analysis)**
  - Classification:
    - Based on existing attribute values, Predict *Nominal/Rank* class labels
    - E.g., "*based on your assignment marks, predict whether you will pass this unit or not*"
      - Output: *Pass*, or *Fail*
      - Or, predict your grade: *F*, *P*, *C*, *D*, *HD*
  - Regression:
    - Based on existing attribute values, Models continuous valued functions, and predict *numerical* values
    - E.g., "*based on your assignment marks, predicate the mark you can get from the examination of this unit*"
      - Output: *real* value, from 0-100

17

Supervised vs Unsupervised Machine Learning

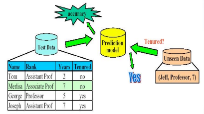
- Supervised Learning, aka **Prediction**
  - Step 1: *Construct a model* to describe the ***training set***
    - The set of data instances used for model construction is called **training data set**
    - ***Data instances*** are also called samples, examples, or items, records, tuples, etc.



18

Supervised vs Unsupervised Machine Learning

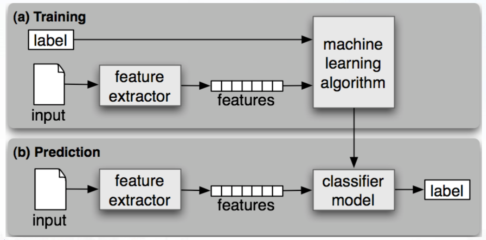
- Supervised Learning, aka **Prediction**
  - Step 2: Use the model to predict *unseen* instances
  - Before use the model, we can estimate the **accuracy** of the model by a ***test data set***
    - Test data set is independent of training data set, and the expected output of test instance is compared with the actual output from the model
    - For *classification*, the accuracy is usually measured by the *percentage of test instances that are correctly classified by the model*
    - For *regression* estimation, the accuracy is usually measured by *mean squared error*



19

Supervised vs Unsupervised Machine Learning

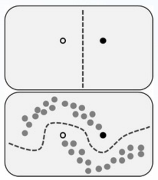
- Flow of Supervised Learning



20

Supervised vs Unsupervised Machine Learning

- Semi-Supervised Learning
  - The top panel shows a decision boundary we might adopt.
  - The bottom panel shows a decision boundary we might adopt if, in addition to the two labeled examples, we were given a collection of unlabeled data.
  - This could be viewed as performing clustering and then labeling the clusters with the labeled data, learning an underlying one-dimensional manifold where the data reside.



21

Machine Learning in Python

- scikit-learn package
  - Machine learning toolboxes in python.
  - Easy to use, plenty of examples and demo.
  - URL: http://scikit-learn.org/
- Machine Learning Library (MLlib)
  - Machine learning toolboxes in Spark.
  - URL: https://spark.apache.org/docs/latest/ml-guide.html

22

Naïve Bayes Classifier (NBC)

- NBC Training
- Prediction with NBC

23

Play Football: Training Data Set

- Consider a decision to *Play* outdoor or not [Weka, ch4]
  - Suppose we've collected the data for the past 2 weeks

Day	Outlook	Temperature	Humidity	Wind	Play Football
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Weak	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

24

Play Football: Training Data Set

- For a new day with following measures for Outlook, Temperature, Humidity and Windy, what is the prediction for Play?

Day	Outlook	Temperature	Humidity	Wind	Play Football
D15	Overcast	Hot	Normal	Weak	????

  - We can answer question like this with many different methods
    - Naïve Bayes Classifier (NBC)
    - or Decision Trees

25

Meet the NBC

- It is a **Bayesian** method
  - A simple but very effective classification method
  - Training data can be discarded once trained.
  - Follows a typical setting for a supervised learning method in machine learning
    - Step 1: train the model using training data (e.g., MLE)
    - Step 2: use trained model to make prediction

26

Train an NBC

	This is the feature X_1	This is the feature X_2	This is the feature X_3	This is the feature X_4	This is the class label Y
Day	Outlook	Temperature	Humidity	Wind	Play Football
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Weak	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

27

Train an NBC

- We need introduce three concepts:
  - Marginal probability
  - Joint probability
  - and Conditional probability

28

Train an NBC

- Marginal probability
  - What is the probability Pr(Outlook = sunny)?

Outlook
Sunny
Sunny
Overcast
Rain
Rain
Rain
Overcast
Sunny
Sunny
Rain
Sunny
Overcast
Overcast
Rain

Method 1: use the frequency table

Outlook	Frequency
sunny	5
overcast	4
rain	5

$$\Pr(\text{Outlook=sunny}) = \frac{5}{14}$$

$$\Pr(\text{Outlook=overcast}) = \frac{4}{14}$$

$$\Pr(\text{Outlook=rain}) = \frac{5}{14}$$

Outlook	Probability
sunny	5/14
overcast	4/14
rain	5/14

Marginal probability for Outlook

29

Train an NBC

- Marginal probability
  - What is the probability Pr(Play = yes)?
    - Method 1: use the frequency table

Play Football
No
No
Yes
Yes
Yes
No
Yes
No
Yes
Yes
Yes
Yes
Yes
No

Play	Frequency
yes	9
no	5

$\Pr(\text{Play}=\text{yes}) = \frac{9}{14}$

$\Pr(\text{Play}=\text{no}) = \frac{5}{14}$

Play	Probability
yes	9/14
no	5/14

Marginal probability for Play

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 30

30

Train an NBC

- Joint probability Pr(Outlook = sunny, Play = yes)
  - Construct the contingency table for Outlook and Play

Outlook	Play Football
Sunny	No
Sunny	No
Overcast	Yes
Rain	Yes
Rain	Yes
Rain	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rain	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rain	No

Outlook \ Play	yes	no
sunny	2	3
overcast	4	0
rainy	3	2

$\Pr(\text{Outlook} = \text{sunny}, \text{Play} = \text{yes}) = \frac{2}{14}$

$\Pr(\text{Outlook} = \text{sunny}, \text{Play} = \text{no}) = \frac{3}{14}$

$\Pr(\text{Outlook} = \text{overcast}, \text{Play} = \text{yes}) = \frac{4}{14}$

....

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 31

31

Train an NBC

- Joint probability Pr(Outlook = sunny, Play = yes)
  - Construct the contingency table for Outlook and Play

Outlook	Play Football
Sunny	No
Sunny	No
Overcast	Yes
Rain	Yes
Rain	Yes
Rain	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rain	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rain	No

Outlook \ Play	yes	no
sunny	2	3
overcast	4	0
rainy	3	2

Contingency table

Outlook \ Play	yes	no
sunny	2/14	3/14
overcast	4/14	0
rainy	3/14	2/14

Joint probability distribution for Outlook and Play

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 32

32

Train an NBC

- Marginal probability Pr(Outlook = sunny)
  - the marginalization rule: $\Pr(X) = \sum_y P(X, Y = y)$

Outlook	Play Football
Sunny	No
Sunny	No
Overcast	Yes
Rain	Yes
Rain	Yes
Rain	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rain	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rain	No

$\Pr(\text{Outlook}=\text{rainy}) = \frac{3}{14} + \frac{2}{14} = \frac{5}{14}$

$\Pr(\text{Outlook}=\text{sunny})$

$= \Pr(\text{Outlook}=\text{sunny}, \text{Play}=\text{yes}) + \Pr(\text{Outlook}=\text{sunny}, \text{Play}=\text{no})$

$= \frac{2}{14} + \frac{3}{14} = \frac{5}{14}$

Outlook \ Play	yes	no
sunny	2/14	3/14
overcast	4/14	0
rainy	3/14	2/14

Outlook	Probability
sunny	5/14
overcast	4/14
rain	5/14

Marginal probability for Outlook computed before

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 33

33

Train an NBC

- Conditional probability Pr(Outlook=sunny | Play = yes)
  - Construct the contingency table for Outlook and Play

Outlook	Play Football
Sunny	No
Sunny	No
Overcast	Yes
Rain	Yes
Rain	Yes
Rain	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rain	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rain	No

Outlook \ Play	yes	no
sunny	2	3
overcast	4	0
rainy	3	2

$\Pr(\text{Outlook} = \text{sunny} \mid \text{Play} = \text{yes}) = \frac{2}{9}$

$\Pr(\text{Outlook} = \text{overcast} \mid \text{Play} = \text{yes}) = \frac{4}{9}$

$\Pr(\text{Outlook} = \text{rainy} \mid \text{Play} = \text{yes}) = \frac{3}{9}$

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 34

34

Train an NBC

- Conditional probability Pr(Outlook=sunny | Play = yes)
  - Construct the contingency table for Outlook and Play

Outlook	Play Football
Sunny	No
Sunny	No
Overcast	Yes
Rain	Yes
Rain	Yes
Rain	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rain	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rain	No

Outlook \ Play	yes	no
sunny	2	3
overcast	4	0
rainy	3	2

$\Pr(\text{Outlook} = \text{sunny} \mid \text{Play} = \text{no}) = \frac{3}{5}$

$\Pr(\text{Outlook} = \text{overcast} \mid \text{Play} = \text{no}) = 0$

$\Pr(\text{Outlook} = \text{rainy} \mid \text{Play} = \text{no}) = \frac{2}{5}$

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 35

35

Train an NBC

- Conditional probability Pr(Outlook=sunny | Play = yes)
  - Construct the contingency table for Outlook and Play

Outlook	Play Football
Sunny	No
Sunny	No
Overcast	Yes
Rain	Yes
Rain	Yes
Rain	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rain	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rain	No

Play

Outlook	yes	no
sunny	2	3
overcast	4	0
rainy	3	2

$\Pr(\text{Outlook} = \text{sunny} \mid \text{Play} = \text{no}) = \frac{3}{5}$

$\Pr(\text{Outlook} = \text{overcast} \mid \text{Play} = \text{no}) = 0$

$\Pr(\text{Outlook} = \text{rainy} \mid \text{Play} = \text{no}) = \frac{2}{5}$

Play

Outlook	yes	no
sunny	2/9	3/5
overcast	4/9	0
rainy	3/9	2/5

conditional

36

Train an NBC

- Training = conditional distribution for class Play

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	Str.	2/9	3/5	9/14	5/14
overcast	4/9	0	mild	4/9	2/5	normal	6/9	1/5	weak	7/9	2/5		
rainy	3/9	2/5	cold	3/9	1/5								

Marginal or prior for Play

37

Prediction with NBC

- Predict Play outcome for a new day

Day	Outlook	Temperature	Humidity	Wind	Play Football
D15	Overcast	Hot	Normal	Weak	????

- This amounts to compute the conditional probability:
  - $\Pr(P = yes | O = overcast, T = hot, H = normal, W = weak)$

  versus
  - $\Pr(P = no | O = overcast, T = hot, H = normal, W = weak)$
- If the probability for P=yes > P=no, then predict yes, else predict no.

38

Prediction with NBC

- Predict Play outcome for a new day

Day	Outlook	Temperature	Humidity	Wind	Play Football
D15	Overcast	Hot	Normal	Weak	????

- How to calculate the class conditional probability?
  - $\Pr(P = yes | O = overcast, T = hot, H = normal, W = weak)$
- We need the Bayes rule

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

constant w.r.t. θ

$$\underbrace{p(\theta \mid D)}_{\textit{posterior}} \propto \underbrace{p(\theta)}_{\textit{prior}} \times \underbrace{p(D \mid \theta)}_{\textit{likelihood}}$$

39

Prediction with NBC

- Predict Play outcome for a new day

Day	Outlook	Temperature	Humidity	Wind	Play Football
D15	Overcast	Hot	Normal	Weak	????

- How to calculate the class conditional probability?

$\Pr(P = yes | O = overcast, T = hot, H = normal, W = weak)$

$\propto \Pr(O = overcast, T = hot, H = normal, W = no \mid P = yes) \Pr(P = yes)$

$$\propto \underbrace{\Pr(O = overcast \mid P = yes) \Pr(T = hot \mid P = yes) \Pr(H = normal \mid P = yes) \Pr(W = no \mid P = yes) \Pr(P = yes)}_{\text{Independently factorized = Naïve assumption}}$$

Bayes'rule + Naïve assumption = Naïve Bayes Model

40

Prediction with NBC

- Predict Play outcome for a new day

Day	Outlook	Temperature	Humidity	Wind	Play Football
D15	Overcast	Hot	Normal	Weak	????

- How to calculate the class conditional probability?

$\Pr(P = yes | O = overcast, T = hot, H = normal, W = weak)$

$\propto \Pr(O = overcast, T = hot, H = normal, W = no \mid P = yes) \Pr(P = yes)$

$\propto \Pr(O = overcast \mid P = yes) \Pr(T = hot \mid P = yes) \Pr(H = normal \mid P = yes) \Pr(W = no \mid P = yes) \Pr(P = yes)$

$\propto \frac{4}{9}\frac{2}{9}\frac{6}{9}\frac{7}{9}\frac{9}{14}$

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	$\boxed{2/9}$	2/5	high	3/9	4/5	Str.	2/9	3/5	$\boxed{9/14}$	5/14
overcast	$\boxed{4/9}$	0	mild	4/9	2/5	normal	$\boxed{6/9}$	1/5	weak	$\boxed{7/9}$	2/5		
rainy	3/9	2/5	cold	3/9	1/5								

41

Prediction with NBC

- Predict Play outcome for a new day

Day	Outlook	Temperature	Humidity	Wind	Play Football
D15	Overcast	Hot	Normal	Weak	????

- How to calculate the class conditional probability?

$\Pr(P = yes|O = overcast, T = hot, H = normal, W = weak)$

$\Pr(P = no|O = overcast, T = hot, H = normal, W = weak) \propto 0$

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	Str.	2/9	3/5	9/14	5/14
overcast	4/9	0	mild	4/9	2/5	normal	6/9	1/5	weak	7/9	2/5		
rainy	3/9	2/5	cold	3/9	1/5								

42

Prediction with NBC

- When data is insufficient, NBC may 'overfit' what it sees
- This probability will always zero:

$\Pr(P = no|O = overcast, T = ?, H = ?, W = ?) = 0$

  – Why? because $\Pr(H = no|O = overcast) = 0$
- How to resolve this?
  – Answer: add a pseudo-count, e.g., 1 to every entry.

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	Str.	2/9	3/5	9/14	5/14
overcast	4/9	0	mild	4/9	2/5	normal	6/9	1/5	weak	7/9	2/5		
rainy	3/9	2/5	cold	3/9	1/5								

43

Prediction with NBC

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	3/12	4/8	hot	3/12	3/8	high	4/11	5/7	Str.	3/11	4/7	10/16	6/16
overcast	5/12	1/8	mild	5/12	3/8	normal	7/11	2/7	weak	8/11	3/7		
rainy	4/12	3/8	cold	4/12	2/8								

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	Str.	2/9	3/5	9/14	5/14
overcast	4/9	0	mild	4/9	2/5	normal	6/9	1/5	weak	7/9	2/5		
rainy	3/9	2/5	cold	3/9	1/5								

44

Prediction with NBC

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	3/12	4/8	hot	3/12	3/8	high	4/11	5/7	Str.	3/11	4/7	10/16	6/16
overcast	5/12	1/8	mild	5/12	3/8	normal	7/11	2/7	weak	8/11	3/7		
rainy	4/12	3/8	cold	4/12	2/8								

After adding pseudo-count to avoid overfitting, what is the prediction for Play now?

Day	Outlook	Temperature	Humidity	Wind	Play Football
D15	Overcast	Hot	Normal	Weak	????

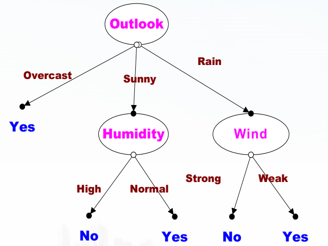
45

Decision Trees

- What is it?
- History of Decision Tree research

46

Decision Tree Representation



A Tree in which

- Each **internal** node is a *test of an attribute*
- Each test has mutually *exclusive and exhaustive* outcomes
- Each **branch** corresponds to an attribute value
- Each **leaf** node assigns a decision

47

Advantages of Decision Trees

- Natural and succinct, suitable for ***Classification*** Problems
  - Classify an example into one of a discrete set of possible values
- If *decision trees* can be *built* automatically from *training data set*, it can be used as a kind of Knowledge Discovery method.

How to Build a Decision Tree from Training Data Set AUTOMATICALLY?

48

Brief history of Decision Tree Construction

- The first decision tree algorithm is ***CLS*** (*Concept Learning System*)
  - E.B.Hunt, J.Martin, and P.T.Stone's book published by Academic Press in 1966
- The algorithm raising the interests in Decision Tree is ***ID3***
  - J.R. Quinlan's paper in a book edited by D. Michie, published by Gordon and Breach in 1979
- The most popular decision tree algorithm that can be used in regression is ***CART*** (*Classification and Regression Tree*)
  - L.Breiman, J.H.Friedman, R.A.Olshen, and C.J.Stone's book published by Wadsworth in 1984
- The current decision tree algorithms include ***C4.5*** (and ***C5***)
  - J.R.Quinlan's book published by Morgan Kaufmann in 1993

49

Major Decision Tree Algorithms

- ID3
  - by ***Ross Quinlan*** 79
  - Uses "***Information Gain***" to select the attributes
- C4.5/C5
  - by ***Ross Quinlan*** 93/97
  - Uses "***Gain Ratio***" to select attribute
- CART
  - by ***Brieman*** 84
  - Uses "***Gini Index***" to select attribute

50

Play Football: Training Data Set

- Consider a decision to *Play* outdoor or not [Weka, ch4]
  - Suppose we've collected the data for the past 2 weeks

Day	Outlook	Temperature	Humidity	Wind	Play Football
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Weak	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

51

Information Gain

- Bits
- Entropy
- Conditional Entropy
- Information Gain

52

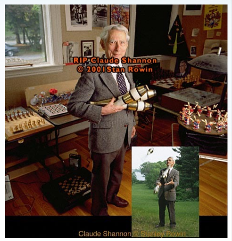
*Claude Shannon

"Father of information theory"

Claude Shannon, who has died aged 84, perhaps more than anyone laid the groundwork for today's digital revolution. His exposition of information theory, stating that all information could be represented mathematically as a succession of noughts and ones, facilitated the digital manipulation of data without which today's information society would be unthinkable.

Shannon's master's thesis, obtained in 1940 at MIT, demonstrated that problem solving could be achieved by manipulating the symbols 0 and 1 in a process that could be carried out automatically with electrical circuitry. That dissertation has been hailed as one of the most significant master's theses of the 20th century. Eight years later, Shannon published another landmark paper, *A Mathematical Theory of Communication*, generally taken as his most important scientific contribution.
Shannon applied the same radical approach to cryptography research, in which he later became a consultant to the US government.

Many of Shannon's pioneering insights were developed before they could be applied in practical form. He was truly a remarkable man, yet unknown to most of the world.



Born: 30 April 1916 Died: 23 February 2001

53

Bits

You are watching a set of independent random samples of *X*

You see that *X* has four possible values: *A, B, C, D*

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

So you might see: *BAACBADCDADDDA…*

You transmit data over a binary serial link. You can encode each reading with two bits (e.g. *A = 00, B = 01, C = 10, D = 11*)

0100001001001110110011111100…

54

Fewer Bits

Someone tells you that the probabilities are not equal

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

It's possible…

…to invent a coding for your transmission that only uses *1.75* bits on average per symbol. How?

A	0
B	10
C	110
D	111

55

Fewer Bits

Suppose there are three equally likely values…

P(X=A) = 1/3	P(X=B) = 1/3	P(X=C) = 1/3

Here's a naïve coding, costing 2 bits per symbol

A	00
B	01
C	10

Can you think of a coding that would need only 1.6 bits per symbol on average?
In theory, it can in fact be done with 1.58496 bits per symbol.

56

General Case

Suppose X can have one of *m* values… $V_1, V_2, \ldots V_m$

$P(X=V_1) = p_1$	$P(X=V_2) = p_2$	…	… $= p_m$

What's the sm… … transmit a stream of s…

A histogram of the frequency distribution of values of X would be flat

A histogram of the frequency distribution of values of X would have many lows and one or two highs

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$

$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$

H(X) = The entropy of X

- "High Entropy" means X is from a uniform (boring) distribution
- "Low Entropy" means X is from varied (peaks and valleys) distribution

57

Entropy in a nut-shell

Low Entropy

..the values (locations of soup) will be to the mouth

High Entropy

..the values (locations of soup) unpredictable... almost uniformly sampled throughout our dining room

58

Example on Entropy

Suppose I'm trying to predict output Y and I have input X

X = University Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Let's assume this reflects the true probabilities

E.G. From this data we estimate

- *P(Y = Yes) = 0.5*
- *P(X = Math & Y = No) = 0.25*
- *P(X = Math) = 0.5*
- *P(Y= Yes | X = History) = 0*

Note:

- *H(X) = 1.5*
- *H(Y) = 1*

59

Specific Conditional Entropy

X = College Major
Y = Likes "Gladiator"

Definition of Specific Conditional Entropy:

*H(Y |X=v) = **The entropy of Y among only those records in which X has value v***

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Example:

H(Y |X=Math) = 1
H(Y |X=History) = 0
H(Y |X=CS) = 0

60

Conditional Entropy

X = College Major
Y = Likes "Gladiator"

Definition of Conditional Entropy :

H(Y |X)

= *The average specific conditional entropy of Y*

= *if you choose a record at random what will be the conditional entropy of Y, conditioned on that row's value of X*

= *Expected number of bits to transmit Y if both sides will know the value of X*

= Σ_j *Prob(X=v*$_j$*) H(Y | X = v*$_j$*)*

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

61

Conditional Entropy

X = College Major
Y = Likes "Gladiator"

Definition of Conditional Entropy:

H(Y |X)

= The average conditional entropy of *Y*

=Σ_j *Prob(X=v*$_j$*) H(Y | X = v*$_j$*)*

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

v_j	Prob(X=v_j)	H(Y | X = v_j)
Math	0.5	1
History	0.25	0
CS	0.25	0

*H(Y|X) = 0.5 * 1 + 0.25 * 0 + 0.25 * 0 = 0.5*

62

Information Gain

X = College Major
Y = Likes "Gladiator"

Definition of Information Gain :

IG(Y|X)
= *I must transmit Y. How many bits on average would it save me if both ends of the line knew X?*
= H(Y) - H(Y | X)

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Example:

- **H(Y) = 1**
- **H(Y|X) = 0.5**
- **Thus IG(Y|X) = 1 – 0.5 = 0.5**

63

Information Gain Example

wealth values: poor rich

gender Female 14423 1769 H(wealth | gender = Female) = 0.497654
Male 22732 9918 H(wealth | gender = Male) = 0.885847

H(wealth) = 0.793844 H(wealth|gender) = 0.757154
IG(wealth|gender) = 0.0366896

64

Another example

wealth values: poor rich

agegroup 10s 2507 3 H(wealth | agegroup = 10s) = 0.0133271
20s 11262 743 H(wealth | agegroup = 20s) = 0.334906
30s 9468 3461 H(wealth | agegroup = 30s) = 0.838134
40s 6738 3986 H(wealth | agegroup = 40s) = 0.951961
50s 4110 2509 H(wealth | agegroup = 50s) = 0.957376
60s 2245 809 H(wealth | agegroup = 60s) = 0.834049
70s 668 147 H(wealth | agegroup = 70s) = 0.680882
80s 115 16 H(wealth | agegroup = 80s) = 0.535474
90s 42 13 H(wealth | agegroup = 90s) = 0.788941

H(wealth) = 0.793844 H(wealth|agegroup) = 0.709463
IG(wealth|agegroup) = 0.0843813

65

Decision Tree Algorithm

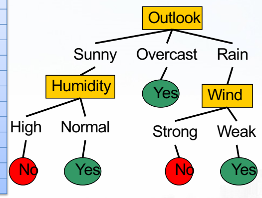
- Decision Tree Algorithm
- Example
- From Tree to Rules
- Evaluation of Decision Tree

66

Classification

- A decision tree for *PlayFootball*

Day	Outlook	Temp	Humid	Wind	PlayFootball
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

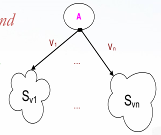


67

ID3: Learning of Decision Trees

- ID3, Concept Learning System(CLS) algorithm
  - *Create a root node for the tree, corresponding to all data examples S*
  - IF *all examples from S belong to the same class C_j,* THEN *label the root with C_j and return*
  - ELSE
    - *Select an attribute A with values $v_1,\ldots,v_n$, and let the root be an Internal node about A*
    - *Partition the data set S into subset $S_1,\ldots,S_n$ according to the values of attribute A*
    - *Apply the algorithm recursively to each subset $S_1,\ldots,S_n$*

How to Select the best attribute?



68

ID3: Search Heuristics

- Which is the attribute that is ***most useful*** for classifying examples?
- **Information Gain**
  - How many ***information*** contained by *Test of an attribute*
    - The larger the Information Gain , the more informative the attribute
- Example:
  - Guess ***Sam***'s gender?
  - If I tell you:
    - Sam's father is a teacher (any hint on Sam's gender?)
    - Sam's mother is a nurse (any hint on Sam's gender?)
    - Sam's husband is a doctor (yay!)

69

ID3: When to Stop?

- Stopping Criterion
  - *If* ***all examples*** *are classified* ***perfectly***, OR
  - ***All attributes*** *are* ***used***
    - *Label the leaf with the most possible class value in the sub training data set.*

70

Example: From Data to Decision Tree

Day	Outlook	Temperature	Humidity	Wind	Play Football
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Weak	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

71

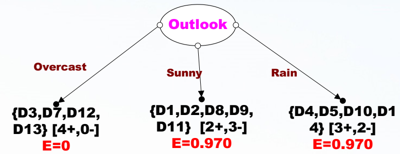
Example

- S={D$_1$,D$_2$,…,D$_{14}$}, written as [9+,5-]
- Class: {Yes, No} for ***PlayFootball***
  - Entropy of S:
    E(S)
    =-(9/14)log$_2$(9/14)-(5/14)log$_2$(5/14)=0.940
  - When calculating entropy, we can use logarithm *based on 2*, or *based on e*, or even *based on 10*. It doesn't affect on our selection of root node.
- Which Attribute as Root of the Tree?
  - Compare the ***Information gain*** of each attribute

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 72

72

Example

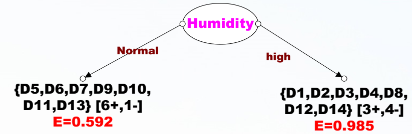


Gain(S,Outlook)
=E(S)-(4/14)*0-(5/14)*0.970-(5/14)*0.970
=0.246

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 73

73

Example



Gain(S,Humidity)
=E(S)-(7/14)*0.985-(7/14)*0.592
=0.151

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 74

74

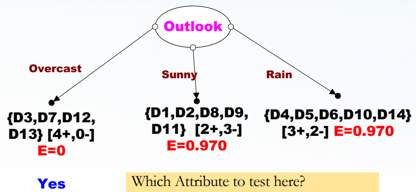
Example

- Gain(S,Wind)=0.102
- Gain(S,Temperature)=0.029
- Gain(S,Outlook)=0.246
- Gain(S,Humidity)=0.151

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 75

75

Example

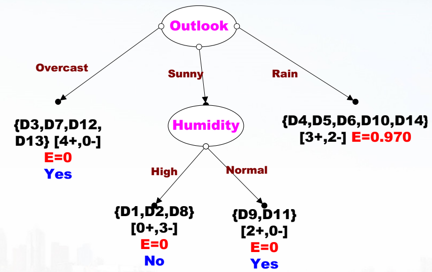


Which Attribute to test here?
- Gain(S$_{sunny}$, Humidity)=0.970
- Gain(S$_{sunny}$, Temperature)=0.570
- Gain(S$_{sunny}$, Wind)=0.019

TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 76

76

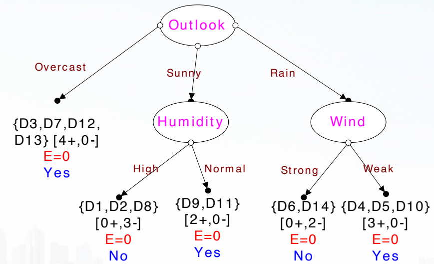
Example



TULIP Team for Universal Learning and Intelligent Processing FLIP: Pattern Recognition (G. Li) 77

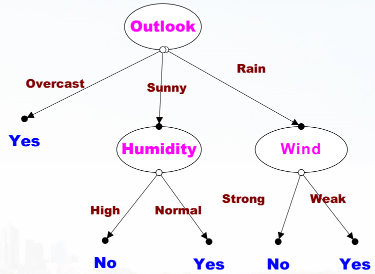
77

Example



78

Example



79

Requirement of ID3

- Discrete Classes (Classification problem)
  - *The values of class should be discrete, such as Yes/No, Young/Old, High/Low, etc*
- Discrete Attributes
  - ID3 also required all attributes to be ***discrete***, otherwise, a ***discretization*** pre-processing step is needed.
- Sufficient Examples
- Complete Data set:
  - *No Missing Value*

80

Other (Im)purity Measures

- Gain Ratio
- Gini Index

81

Highly-branching attributes

- Problematic: ***attributes with a large number of values***
  - extreme case: ID code
- Subsets are more likely to be pure if there is a large number of values
  - *Information gain is biased towards choosing attributes with a large number of values*
  - This may result in ***over fitting***
    - selection of an attribute that is non-optimal for prediction

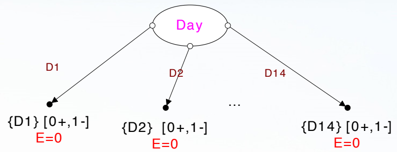
82

Play Football: Training Data Set

Day	Outlook	Temperature	Humidity	Wind	Play Football
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Weak	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

83

Example



Gain(S,Day)
=E(S)-(1/14)*0-(1/14)*0-...-(1/14)*0
=0.940

84

Gain ratio

- ***Gain ratio***: a modification of the information gain that *reduces its bias on high-branch attributes*
- Gain ratio should be
  - *Large when data is evenly spread*
  - *Small when all data belong to one branch*
- Gain ratio takes ***number*** and ***size*** of branches into account when choosing an attribute
  - It corrects the information gain by taking the ***intrinsic information*** of a split into account
    - i.e. how much info do we need to tell which branch an instance belongs to

85

Gain Ratio

- ***Split information***: entropy of distribution of instances into branches

$$SplitInfo(S,A) \equiv -\sum \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}.$$

- ***Gain ratio*** (Quinlan'86) normalizes info gain by:

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInfo(S,A)}.$$

86

Computing the gain ratio

- Example: Split information for "Day"

$$\text{SplitInfo}([1,1,\ldots,1]) = 14 \times (-1/14 \times \log 1/14) = 3.807 \text{ bits}$$

- Example of gain ratio:

$$\text{gain\_ratio}(\text{"Attribute"}) = \frac{\text{gain}(\text{"Attribute"})}{\text{SplitInfo}(\text{"Attribute"})}$$

- Example:

$$\text{gain\_ratio}(\text{"Day"}) = \frac{0.940 \text{ bits}}{3.807 \text{ bits}} = 0.246$$

87

More on the gain ratio

- "Outlook" still comes out top
- However: "Day" still has greater gain ratio
  - Standard fix: ad hoc test to prevent splitting on that type of attribute
- Problem with gain ratio: it may overcompensate
  - May choose an attribute just because its Split information is very low
  - Standard fix:
    - *First, only consider attributes with greater than average information gain*
    - *Then, compare them on gain ratio*

88

Gini Index

- ***Gini Index***: Another sensible measure of impurity (i and j are classes)

$$Gini = \sum_{i \neq j} p(i)p(j)$$

- After applying attribute A, the resulting Gini index is

$$Gini(A) = \sum_{v} p(v) \sum_{i \neq j} p(i|v)p(j|v)$$

- Gini can be interpreted as expected error rate

89

Gini Gain

- The merit of an attribute A can be estimated by ***Gini Gain***, which is

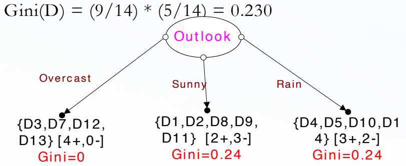
$$GiniGain(A) = Gini - Gini(A)$$

- We can still pick the attribute with the largest Gini Gain in each step

90

Example

- For the PlayFootball Data Set, the original Gini is
  - Gini(D) = (9/14) * (5/14) = 0.230



Gini(Outlook)
=(4/14)*0 + (5/14)*0.24 + (5/14)*0.24=0.171
GiniGain(S,Outlook)=0.230-0.171=0.058

91

Exercise

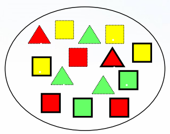
- If you want to master this decision tree learning technique, you are expected to learn decision trees from the given data set:
  - using ***Information Gain*** as a measurement
  - using ***GainRatio*** as a measurement
  - using ***GiniGain*** as a measurement
- then compare what is the difference (if there is) among these three results.

92

Triangles and Squares

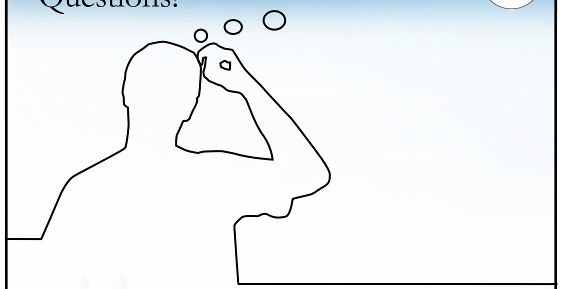
#	Attribute			Shape
	Color	Outline	Dot	
1	green	dashed	no	triange
2	green	dashed	yes	triange
3	yellow	dashed	no	square
4	red	dashed	no	square
5	red	solid	no	square
6	red	solid	yes	triange
7	green	solid	no	square
8	green	dashed	no	triange
9	yellow	solid	yes	square
10	red	solid	no	square
11	green	solid	yes	square
12	yellow	dashed	yes	square
13	yellow	solid	no	square
14	red	dashed	yes	triange

Data Set:
A set of classified objects



93

Questions?



94

This Week's Readings

- Decision Tree
  - K. Murthy, Automatic Construction of Decision Tree from Data: A Multi Disciplinary Survey
- Information Theory
  - C.E. Shannon. *A mathematical theory of Communication.*
- Machine Learning
  - http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- ML Video Series
  - https://www.youtube.com/watch?v=eloiMnin4kk&list=PL5-da3qGB5ICeMbQuqbbCOQWcS6OYBr5A&index=1

95