Machine Learning 10-401, Spring 2018

Introduction, Admin, Course Overview

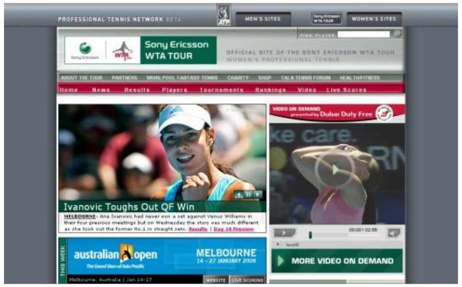
Lecture 1, 01/17/ 2018

Maria-Florina (Nina) Balcan

Machine Learning

Image Classification

Document Categorization



Speech Recognition

Protein Classification

Spam Detection

Branch Prediction

Fraud Detection

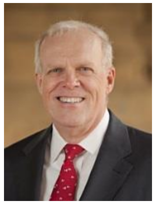
Natural Language Processing

Playing Games

Computational Advertising

Machine Learning is Changing the World

“Machine learning is the hot new thing”
(John Hennessy, President, Stanford)



“A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Microsoft)

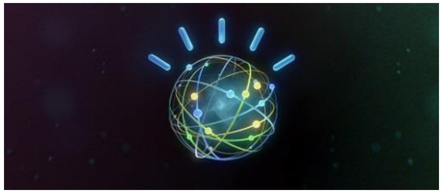
“Web rankings today are mostly a matter of machine learning”
(Prabhakar Raghavan, VP Engineering at Google)

SMARTER THAN YOU THINK

Aiming to Learn as We Do, a Machine Teaches Itself

Jeff Swensen for The New York Times

would sound like me.
So, what you see now is the result of that change.
I'm speaking in english and hopefully you'll hear me speaking chinese in my own voice.
所以，你现在看到的是这一变革的结果。
我是说英语，希望你能听到我说中国人自己的声音。



The COOLEST TOPIC IN SCIENCE

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)
- “Machine learning is the next Internet” (Tony Tether, Director, DARPA)
- Machine learning is the hot new thing” (John Hennessy, President, Stanford)
- “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)
- “Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)
- “Machine learning is today’s discontinuity” (Jerry Yang, CEO, Yahoo)

This course: introduction to machine learning.

- Cover (some of) the most commonly used machine learning paradigms and algorithms.
  - Sufficient amount of details on their mechanisms: explain why they work, not only how to use them.
  - Applications.

What is Machine Learning?

Examples of important machine learning paradigms.

Supervised Classification

from data to discrete classes

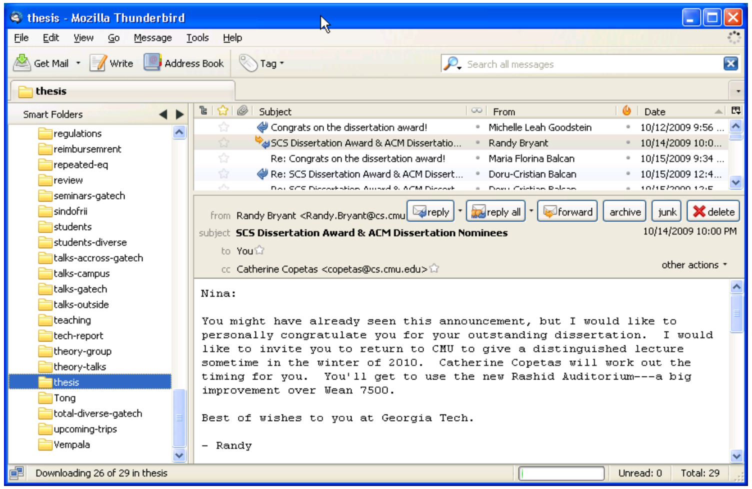
Supervised Classification. Example: Spam Detection

Decide which emails are spam and which are important.

Not spam

Supervised classification

spam



Subject	From	Date
Congrats on the dissertation award!	Michelle Leah Goodstein	10/12/2009 9:56 ...
SCS Dissertation Award & ACM Dissertatio...	Randy Bryant	10/14/2009 10:0...
Re: Congrats on the dissertation award!	Maria Florina Balcan	10/15/2009 9:34 ...
Re: SCS Dissertation Award & ACM Dissert...	Doru-Cristian Balcan	10/15/2009 12:4...

from Randy Bryant <Randy.Bryant@cs.cmu

subject **SCS Dissertation Award & ACM Dissertation Nominees** 10/14/2009 10:00 PM

to You

cc Catherine Copetas <copetas@cs.cmu.edu>

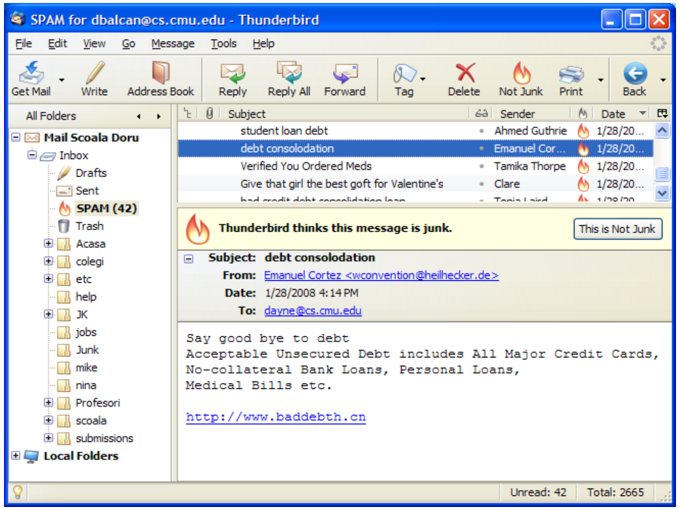
```
Nina:

You might have already seen this announcement, but I would like to
personally congratulate you for your outstanding dissertation.  I would
like to invite you to return to CMU to give a distinguished lecture
sometime in the winter of 2010.  Catherine Copetas will work out the
timing for you.  You'll get to use the new Rashid Auditorium---a big
improvement over Wean 7500.

Best of wishes to you at Georgia Tech.

- Randy
```

Downloading 26 of 29 in thesis | Unread: 0 | Total: 29



Subject	Sender	Date
student loan debt	Ahmed Guthrie	1/28/20...
debt consolodation	Emanuel Cor...	1/28/20...
Verified You Ordered Meds	Tamika Thorpe	1/28/20...
Give that girl the best goft for Valentine's	Clare	1/28/20...

Thunderbird thinks this message is junk.

Subject: debt consolodation
From: Emanuel Cortez <wconvention@heilhecker.de>
Date: 1/28/2008 4:14 PM
To: dayne@cs.cmu.edu

```
Say good bye to debt
Acceptable Unsecured Debt includes All Major Credit Cards,
No-collateral Bank Loans, Personal Loans,
Medical Bills etc.

http://www.baddebth.cn
```

Unread: 42 | Total: 2665

Goal: use emails seen so far to produce good prediction rule for future data.

Supervised Classification. Example: Spam Detection

Represent each message by features. (e.g., keywords, spelling, etc.)

“money”	“pills”	“Mr.”	bad spelling	known-sender	spam?
Y	N	Y	Y	N	Y
N	N	N	Y	Y	N
N	Y	N	N	N	Y
Y	N	N	N	Y	N
N	N	Y	N	Y	N
Y	N	N	Y	N	Y
N	N	Y	N	N	N

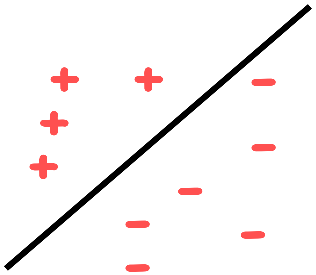
example

label

Reasonable RULES:

Predict SPAM if unknown AND (money OR pills)

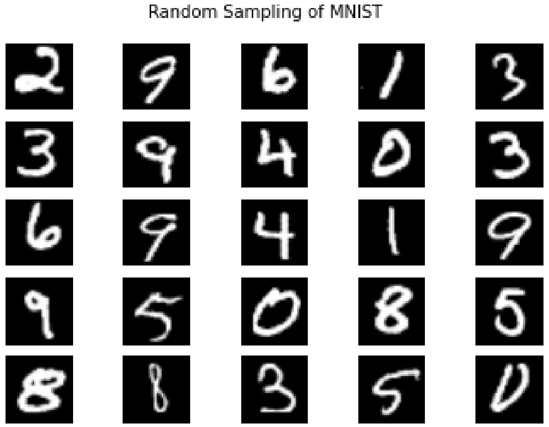
Predict SPAM if 2money + 3pills –5 known > 0



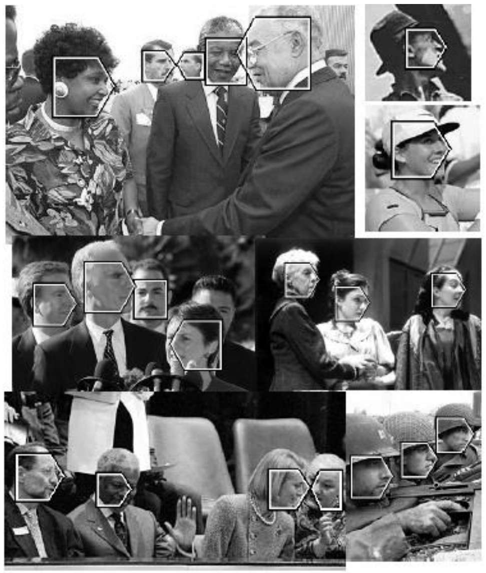
Linearly separable

Supervised Classification. Example: Image classification

- Handwritten digit recognition (convert hand-written digits to characters 0..9)

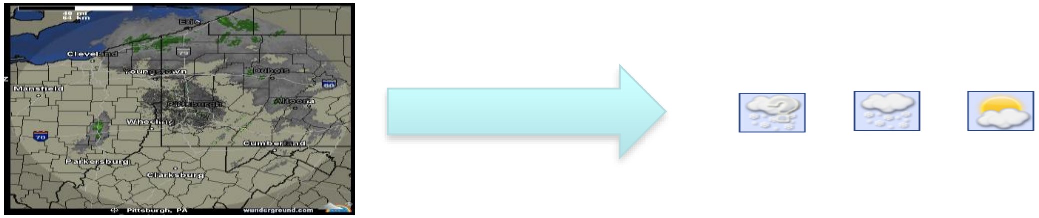


- Face Detection and Recognition



Supervised Classification. Many other examples

- Weather prediction

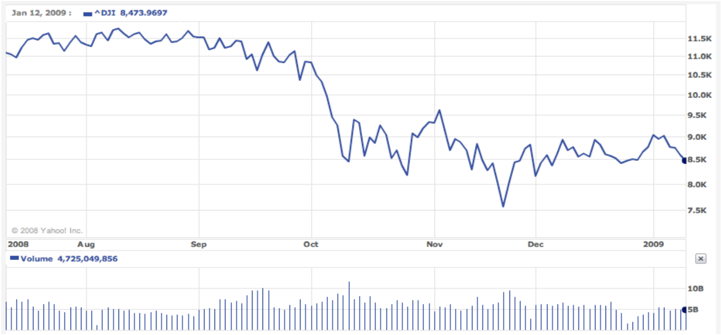


- Medicine:
  - diagnose a disease
    - input: from symptoms, lab measurements, test results, DNA tests, …
    - output: one of set of possible diseases, or “none of the above”
    - examples: audiology, thyroid cancer, diabetes, …
      - or: response to chemo drug X
      - or: will patient be re-admitted soon?

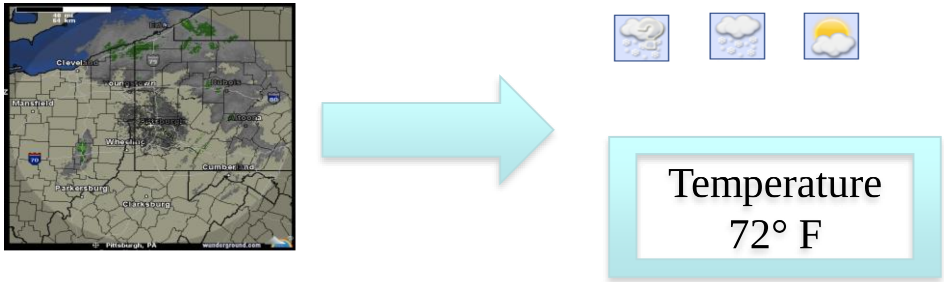
- Computational Economics:
  - predict if a stock will rise or fall
  - predict if a user will click on an ad or not
    - in order to decide which ad to show

Regression. Predicting a numeric value

Stock market



Weather prediction



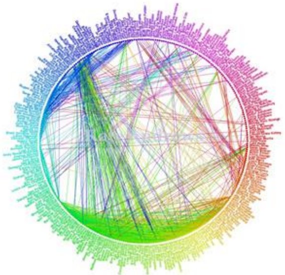
Temperature
72° F

Predict the temperature at any given location

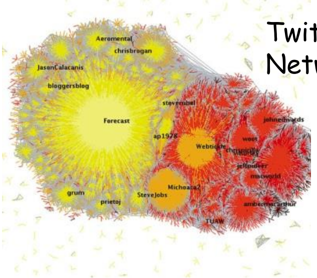
Other Machine Learning Paradigm

Clustering: discovering structure in data (only unlabeled data)

- E.g, cluster users of social networks by interest (community detection).



Facebook network



Twitter Network

Semi-Supervised Learning: learning with labeled & unlabeled data

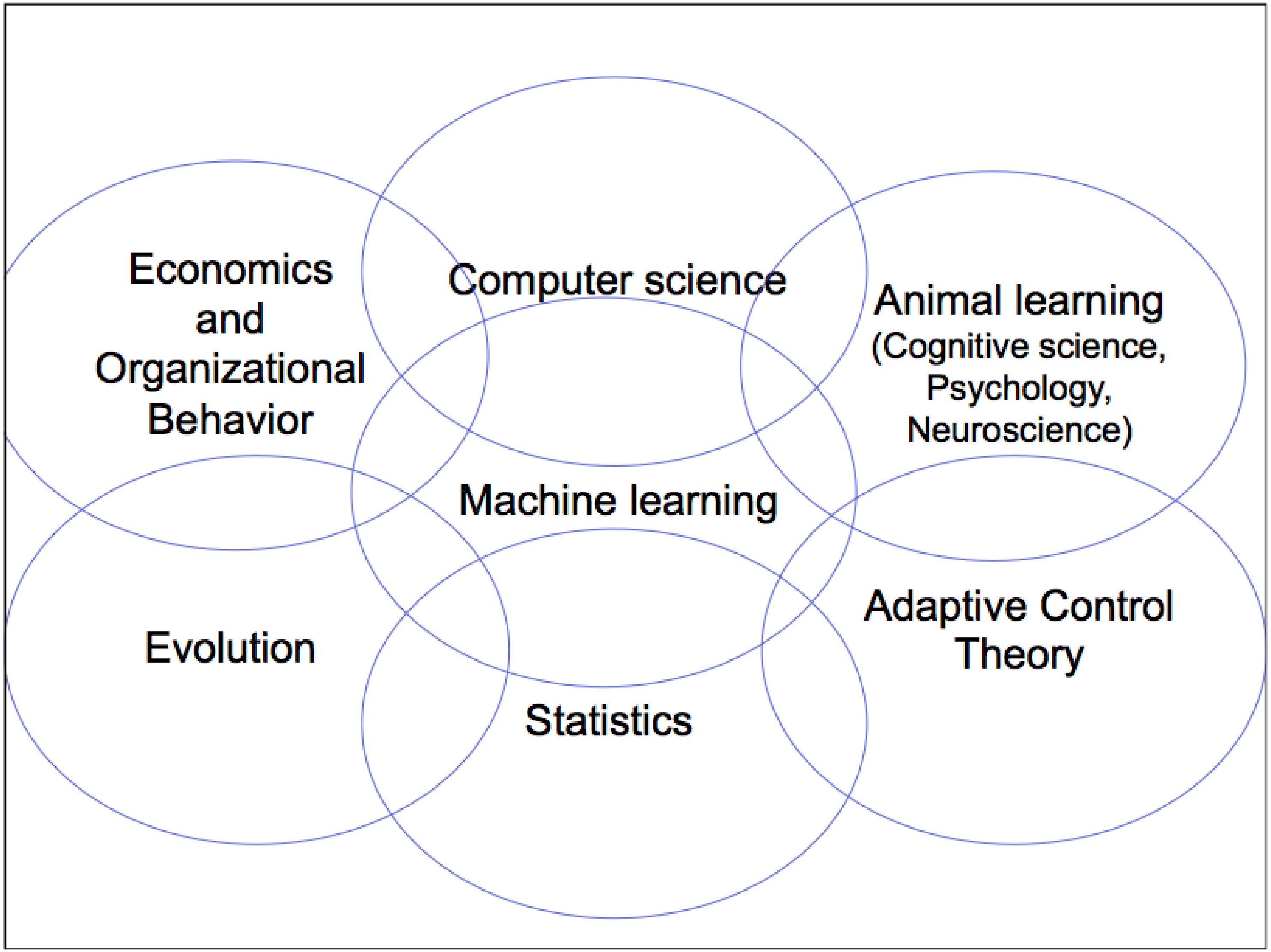
Active Learning: learns pick informative examples to be labeled

Reinforcement Learning (acommodates indirect or delayed feedback)

Dimensionality Reduction

Collaborative Filtering (Matrix Completion), …

Many communities relate to ML


Economics and Organizational Behavior
Computer science
Animal learning (Cognitive science, Psychology, Neuroscience)
Machine learning
Evolution
Statistics
Adaptive Control Theory

Admin, Logistics, Grading

Brief Overview

- Meeting Time: Mon, Wed, NSH 3002, 10:30 – 11:50
- Course Staff
  - Instructors:
    - Maria Florina (Nina) Balcan (ninamf@cs.cmu.edu)
  - TAs:
    - Kenneth Marino (kdmarino@cs.cmu.edu)
    - Colin White (crwhite@cs.cmu.edu)
    - Nupur Chatterji (nchatter@andrew.cmu.edu)

Brief Overview

- Course Website

  http://www.cs.cmu.edu/~ninamf/courses/401sp18

- See website for:
  - Syllabus details
    - All the lecture slides and homeworks
    - Additional useful resources.
  - Office hours
  - Recitation sessions
  - Grading policy
  - Honesty policy
  - Late homework policy
  - Piazza pointers

- Will use Piazza for discussions.

Prerequisites. What do you need to know now?

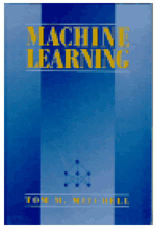
- You should know how to do math and how to program:
  - Calculus (multivariate)
  - Probability/statistics
  - Algorithms. Big O notation.
  - Linear algebra (matrices and vectors)
  - Programming:
    - You will implement some of the algorithms and apply them to datasets
    - Assignments will be in Octave (play with that now if you want; also recitation tomorrow)
    - Octave is open-source software clone of Matlab.
- We may review these things but we will **not** teach them

Source Materials

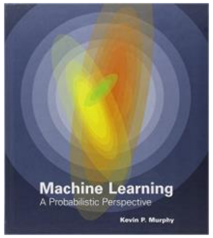
No textbook required. Will point to slides and freely available online material.

Useful textbooks:

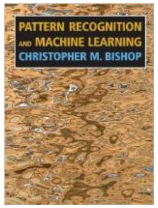
Machine Learning, Tom Mitchell, McGraw Hill, 1997.



Machine Learning: a Probabilistic Perspective, K. Murphy, MIT Press, 2012



Pattern Recognition and Machine Learning
Christopher Bishop, Springer-Verlag 2006



Grading

- 40% for homeworks. There are 5 and you can drop 1.
- 20% for midterm [March 7]
- 20% for final [May 2nd]
- 15% for project
- 5% for class participation.
    - Piazza polls in class: bring a laptop or a phone
- Homework 0: background hwk, out today [get full credit if you turn it in]
- Homeworks 1 to 4
    - Theory/math handouts
    - Programming exercises; applying/evaluating existing learners
    - Late assignments:
        - Up to 50% credit if it’s less than 48 hrs late
        - You can drop your lowest assignment grade
- Projects: conduct a small experiment or read a couple of papers and present the main ideas or work on a small theoretical question.
    - Project presentations: April 23 and April 25

Collaboration policy (see syllabus)

- Discussion of anything is ok…
- …but the goal should be to *understand* better, not save work.

- So:
  - *no notes* of the discussion are allowed…the only thing you can take away is whatever's in your brain.
  - you should acknowledge who you got help from/did help in your homework

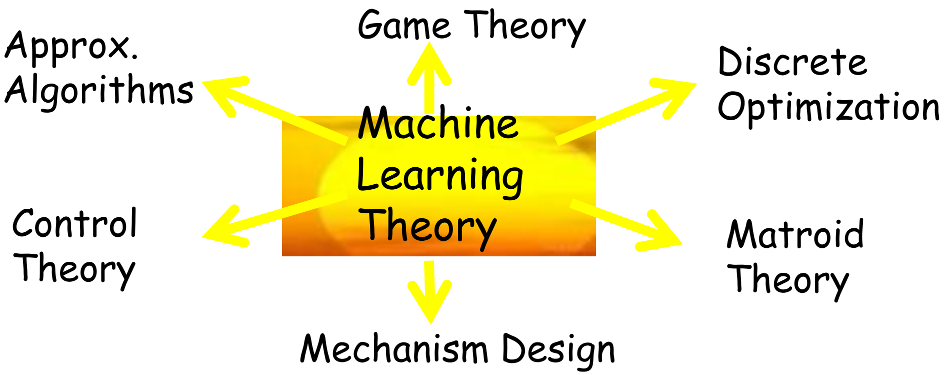
Brief Overview

- Meeting Time: Mon, Wed, NSH 3002, 10:30 – 11:50
- Course Staff
  - Instructors:
    - Maria Florina (Nina) Balcan (ninamf@cs.cmu.edu)
  - TAs:
    - Kenneth Marino (kdmarino@cs.cmu.edu)
    - Colin White (crwhite@cs.cmu.edu)
    - Nupur Chatterji (nchatter@andrew.cmu.edu)

Maria-Florina Balcan: Nina

- Foundations for Modern Machine Learning
  - E.g., interactive, semi-supervised, distributed, life-long learning

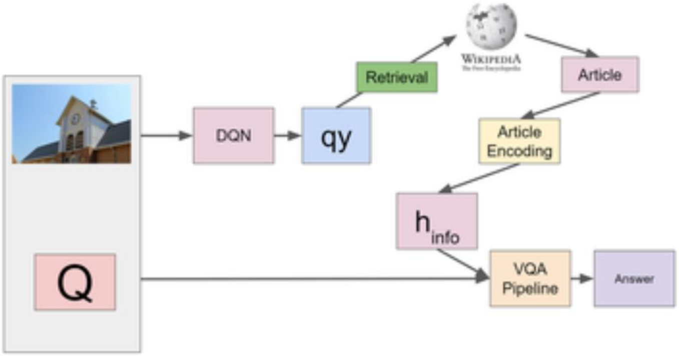
- Connections between learning & other fields (algorithms, algorithmic game theory)



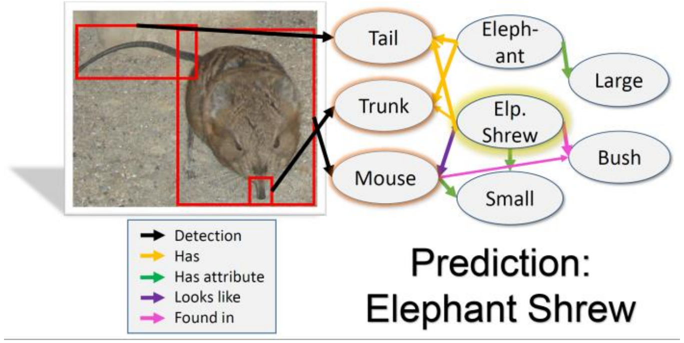
- Program Committee Chair for ICML 2016, COLT 2014

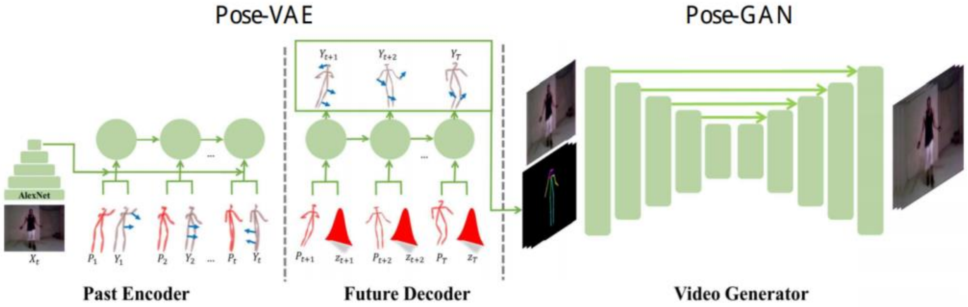
Kenneth Marino: Kenny

- Incorporating knowledge into Computer Vision
  - Incorporating knowledge graphs
  - Learning from Wikipedia articles



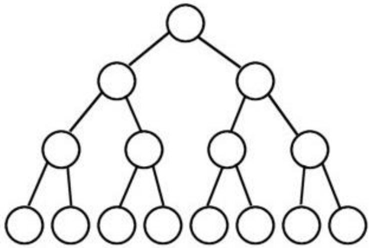
- Deep Learning – non-traditional training and architectures
  - Graph Networks
  - Generative Models (VAEs and GANs)



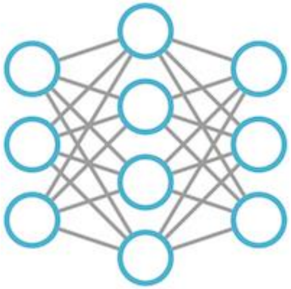


Colin White

- 4th year PhD student advised by Nina Balcan
- Design and analysis of algorithms



- Theoretical foundations of machine learning

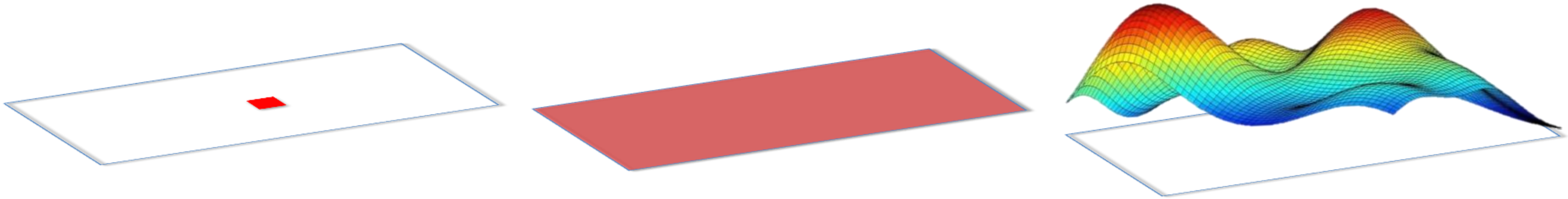


- Beyond worst-case analysis

Worst-case

Average case

Real-world, application-specific



Nupur Chatterji

- Senior is SCS (Undergrad)
- Minor in Machine Learning (and Economics)
- Intend to pursue ML in grad school
- Interested in the intersection between technology and healthcare

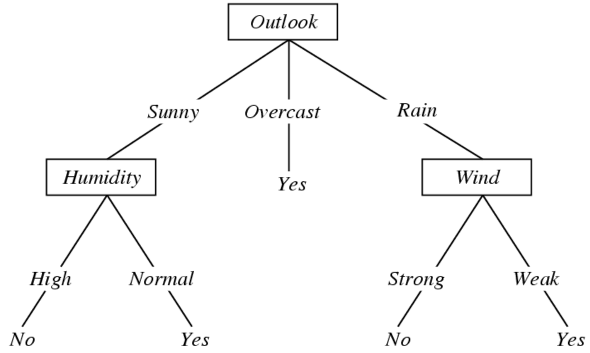
Learning Decision Trees.

Supervised Classification.

Useful Readings:

- Mitchell, Chapter 3
- Bishop, Chapter 14.4

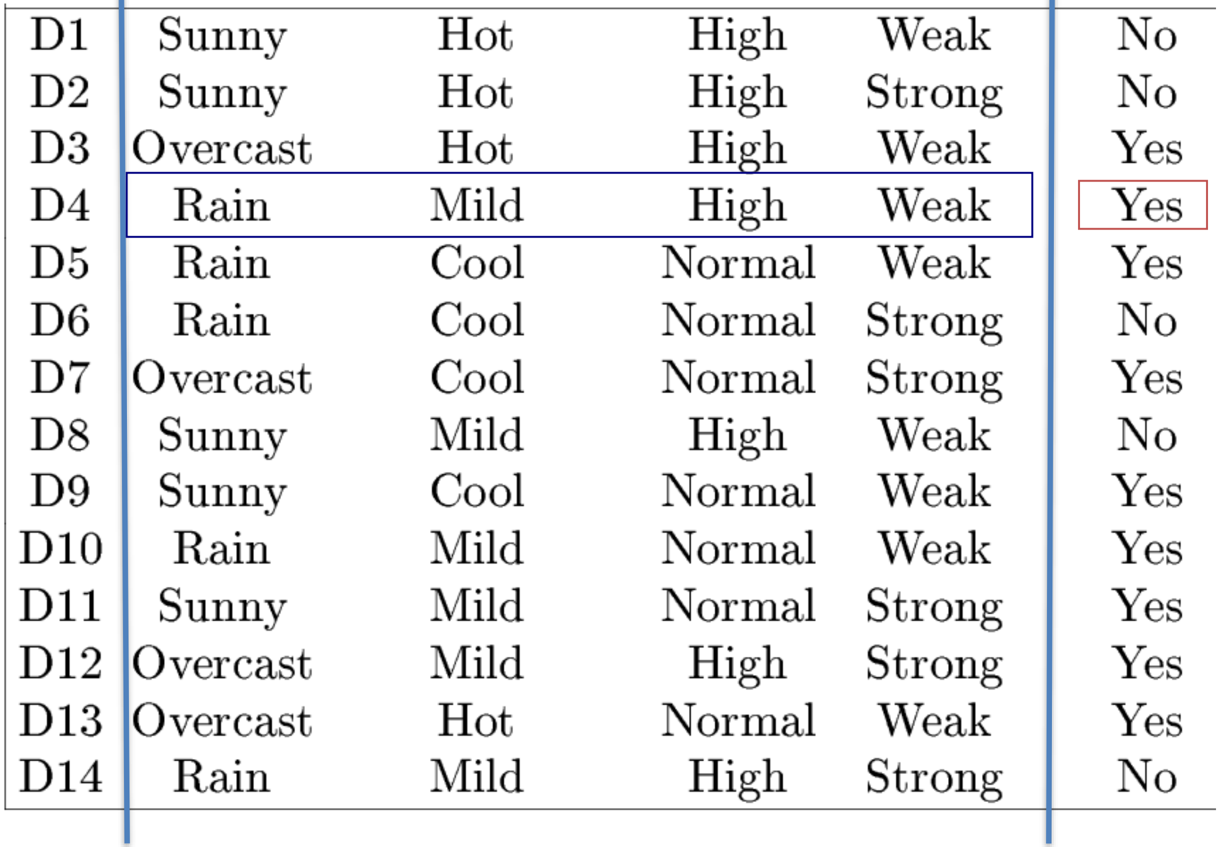
DT learning: Method for learning discrete-valued target functions in which the function to be learned is represented by a decision tree.



Supervised Classification: Decision Tree Learning

Example: learn concept **PlayTennis** (i.e., decide whether our friend will play tennis or not in a given day)

Simple Training Data Set



Day	Outlook	Temperature	Humidity	Wind	Play Tennis
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

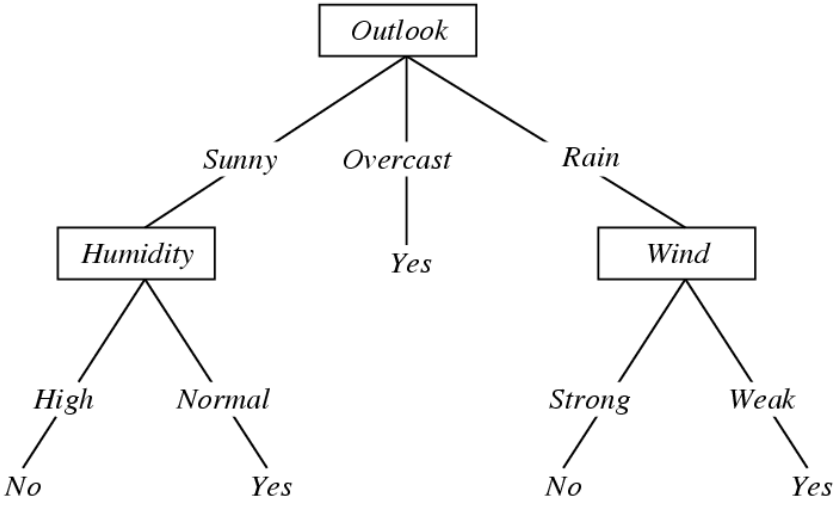
example

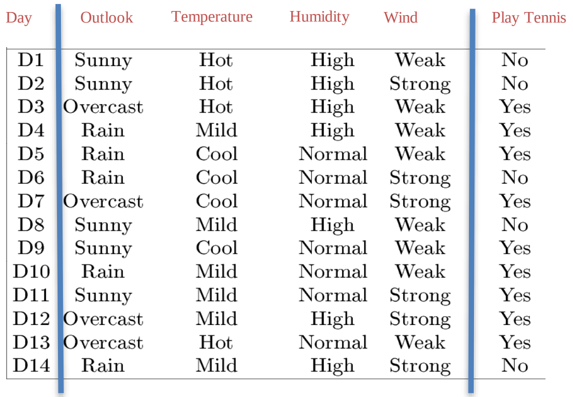
label

Supervised Classification: Decision Tree Learning

- Each internal node: test one (discrete-valued) attribute X_i
- Each branch from a node: corresponds to one possible values for X_i
- Each leaf node: predict Y (or $P(Y=1|x \in \text{leaf})$)

Example: A Decision tree for
f: <Outlook, Temperature, Humidity, Wind> → PlayTennis?





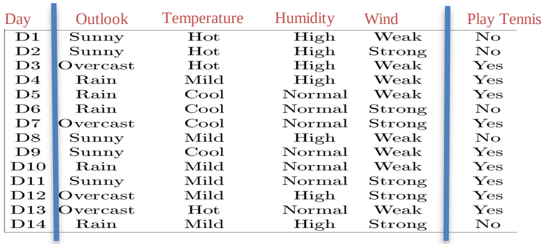
Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	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

E.g., x=(Outlook=sunny, Temperature-Hot, Humidity=Normal,Wind=High), f(x)=Yes.

Supervised Classification: Problem Setting

Input: Training labeled examples $\{(x^{(i)},y^{(i)})\}$ of unknown target function f

- Examples described by their values on some set of features or attributes
  - E.g. 4 attributes: *Humidity*, *Wind*, *Outlook*, *Temp*
    - e.g., <*Humidity=High*, *Wind=weak*, *Outlook=rain*, *Temp=Mild*>
  - Set of possible instances X (a.k.a instance space)

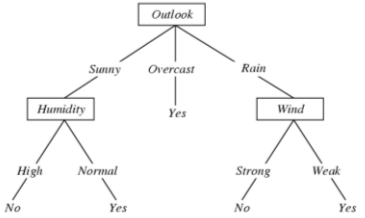


Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	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

- Unknown target function $f : X \rightarrow Y$
  - e.g., $Y=\{0,1\}$ label space
  - e.g., 1 if we play tennis on this day, else 0

Output: Hypothesis $h \in H$ that (best) approximates target function f

- Set of function hypotheses $H=\{ h \mid h : X \rightarrow Y \}$
  - each hypothesis h is a decision tree



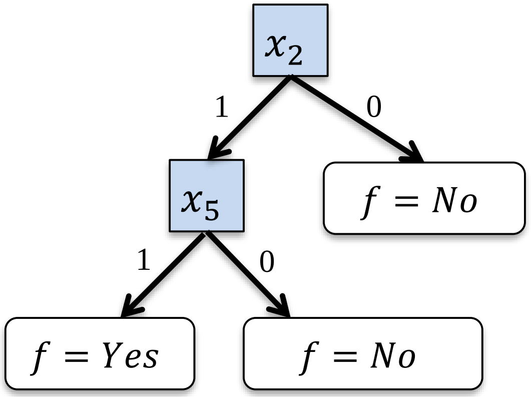
Supervised Classification: Decision Trees

Suppose $X = <x_1, \dots x_n>$

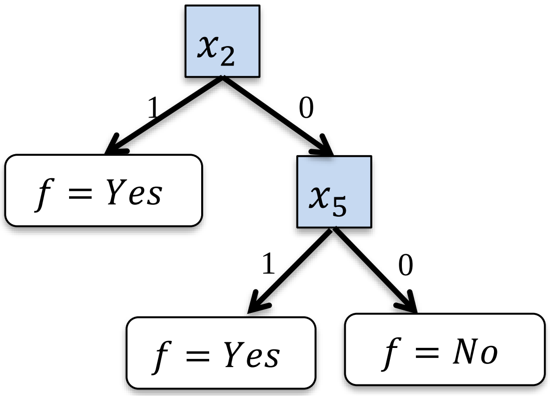
where x_i are boolean-valued variables

How would you represent the following as DTs?

$f(x) = x_2 \;\; AND \; x_5$?



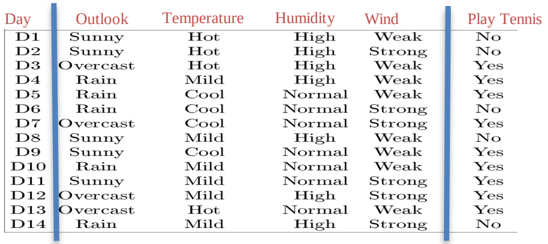
$f(x) = x_2 \; OR \; x_5$



Hwk: How would you represent $X_2 \, X_5 \;\vee\; X_3X_4(\neg X_1)$?

Supervised Classification: Problem Setting

Input: Training labeled examples $\{(x^{(i)},y^{(i)})\}$ of unknown target function f



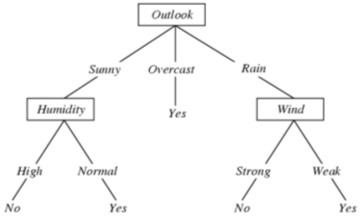
Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	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

- Examples described by their values on some set of features or attributes
  - E.g. 4 attributes: *Humidity*, *Wind*, *Outlook*, *Temp*
    - e.g., <*Humidity*=*High*, *Wind*=*weak*, *Outlook*=*rain*, *Temp*=*Mild*>
  - Set of possible instances X (a.k.a instance space)

- Unknown target function $f : X \rightarrow Y$
  - e.g., Y={0,1} label space
  - e.g., 1 if we play tennis on this day, else 0

Output: Hypothesis $h \in H$ that (best) approximates target function f

- Set of function hypotheses H={ h | $h : X \rightarrow Y$ }
  - each hypothesis h is a decision tree



Core Aspects in Decision Tree & Supervised Learning

How to automatically find a good hypothesis for training data?

- This is an **algorithmic** question, the main topic of computer science

When do we generalize and do well on unseen data?

- **Learning theory** quantifies ability to ***generalize*** as a function of the amount of training data and the hypothesis space
- **Occam's razor:** use the *simplest* hypothesis consistent with data!

Fewer short hypotheses than long ones

- a short hypothesis that fits the data is less likely to be a statistical coincidence
- highly probable that a sufficiently complex hypothesis will fit the data

Core Aspects in Decision Tree & Supervised Learning

How to automatically find a good hypothesis for training data?

- This is an **algorithmic** question, the main topic of computer science

When do we generalize and do well on unseen data?

- **Occam's razor:** use the *simplest* hypothesis consistent with data!

- Decision trees: if we were able to find a small decision tree that explains data well, then good generalization guarantees.
  - NP-hard [Hyafil-Rivest'76]: unlikely to have a poly time algorithm

- Very nice practical heuristics; top down algorithms, e.g, ID3

Top-Down Induction of Decision Trees

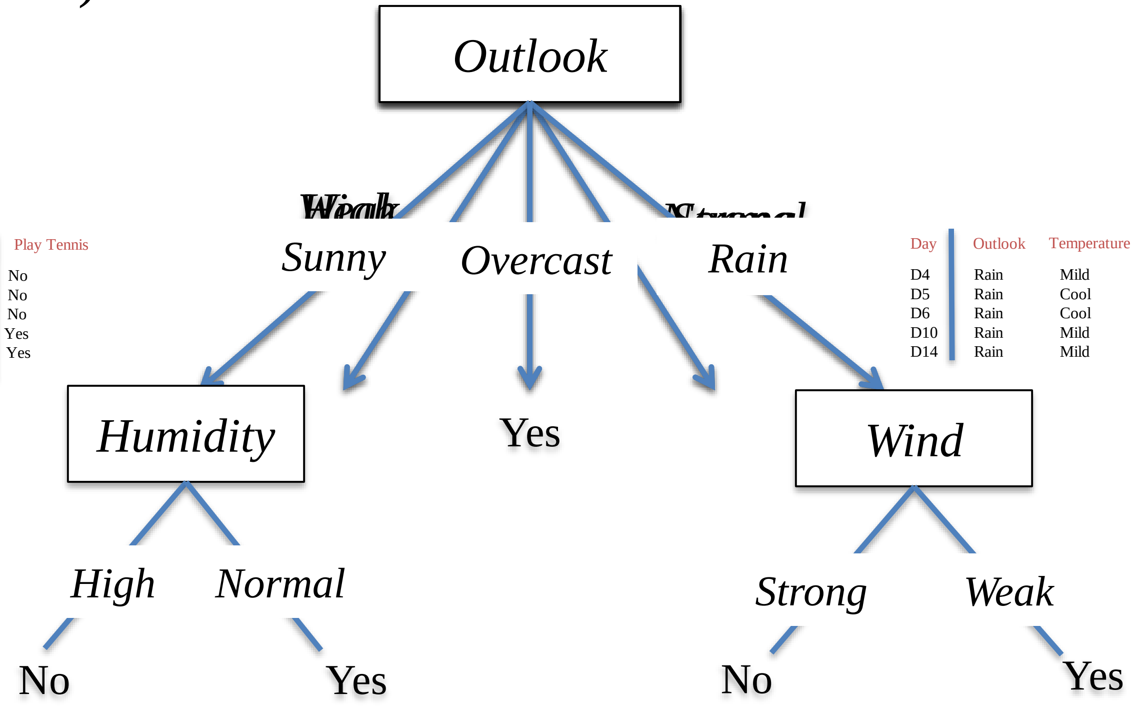
[ID3, C4.5, Quinlan]

ID3: Natural greedy approach to growing a decision tree top-down (from the root to the leaves by repeatedly replacing an existing leaf with an internal node.).

Algorithm:

- Pick “best” attribute to split at the root based on training data.
- Recurse on children that are impure (e.g, have both Yes and No).

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
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



Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D10	Rain	Mild	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No