Introduction to Machine Learning

Machine Learning: Jordan Boyd-Graber
University of Maryland
LECTURE 1A

Slides adapted from Lauren Hannah and Dave Blei

Roadmap

- What machine learning is
- What machine learning can do
- What the course is about

Data are everywhere.

User ratings

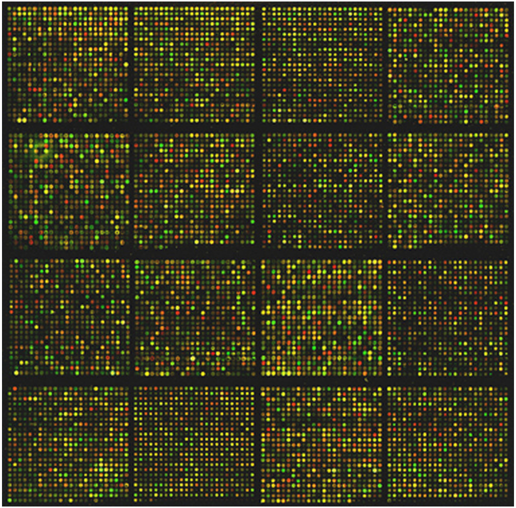
Ikiru (1952)	UR	Foreign
Junebug (2005)	R	Independent
La Cage aux Folles (1979)	R	Comedy
The Life Aquatic with Steve Zissou (2004)	R	Comedy
Lock, Stock and Two Smoking Barrels (1998)	R	Action & Adventure
Lost in Translation (2003)	R	Drama
Love and Death (1975)	PG	Comedy
The Manchurian Candidate (1962)	PG-13	Classics
Memento (2000)	R	Thrillers
Midnight Cowboy (1969)	R	Classics

Purchase histories

	Cheese				
0.5/0.51 lb	**Cabot Vermont Cheddar**	0.51 lb	$7.99/lb	**$4.07**	
	Dairy				
1/1	**Friendship Lowfat Cottage Cheese** (16oz)		$2.89/ea	**$2.89**	
1/1	**Nature's Yoke Grade A Jumbo Brown Eggs** (1 dozen)		$1.49/ea	**$1.49**	
1/1	**Santa Barbara Hot Salsa, Fresh** (16oz)		$2.69/ea	**$2.69**	
1/1	**Stonyfield Farm Organic Lowfat Plain Yogurt** (32oz)		$3.59/ea	**$3.59**	
	Fruit				
3/3	**Anjou Pears** (Farm Fresh, Med)	1.76 lb	$2.49/lb	**$4.38**	
2/2	**Cantaloupe** (Farm Fresh, Med)		$2.00/ea	**$4.00**	**S**
	Grocery				
1/1	**Fantastic World Foods Organic Whole Wheat Couscous** (12oz)		$1.99/ea	**$1.99**	
1/1	**Garden of Eatin' Blue Corn Chips** (9oz)		$2.49/ea	**$2.49**	
1/1	**Goya Low Sodium Chickpeas** (15.5oz)		$0.89/ea	**$0.89**	
2/2	**Marcal 2-Ply Paper Towels, 90ct** (1ea)		$1.09/ea	**$2.18 T**	
1/1	**Muir Glen Organic Tomato Paste** (6oz)		$0.99/ea	**$0.99**	
1/1	**Starkist Solid White Albacore Tuna in Spring Water** (6oz)		$1.89/ea	**$1.89**	

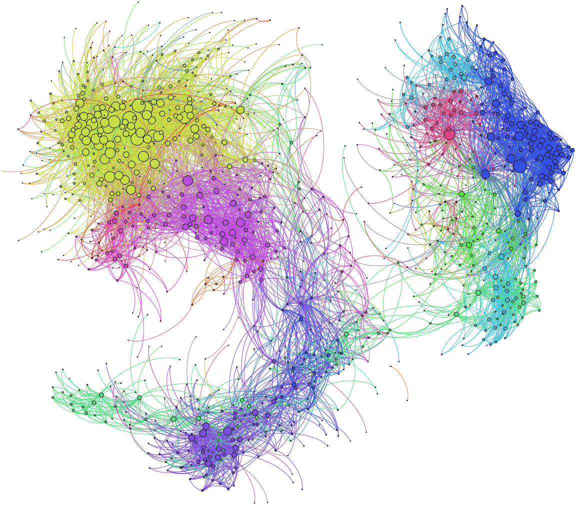
Document collections

Genomics



Neuroscience

Social networks



Finance


Dow Jones Composite Average (^DJA) - DJI
Add to Portfolio
4,372.78 20.68 (0.47%) 1:48PM EDT
Enter name(s) or symbol(s)
GET CHART
COMPARE
EVENTS
TECHNICAL INDICATORS
CHART SETTINGS
RESET
Aug 29, 2012 2:59 PM - 3:04 PM EDT: ^DJA 4416.03
4,420
4,410
4,400
4,390
4,380
4,370
© 2012 Yahoo! Inc.
Tue Aug 28, 2012
Wed Aug 29
Thu Aug 30
Fri Aug 31
Tue Sep 4
Volume: 571,700
8.0M
6.0M
4.0M
2.0M
1D 5D 1M 3M 6M YTD 1Y 2Y 5Y Max
FROM: Aug 28 2012
TO: Sep 4 2012
-0.97%
1980
1985
1990
1995
2000
2005
2010

Data can help us solve problems.

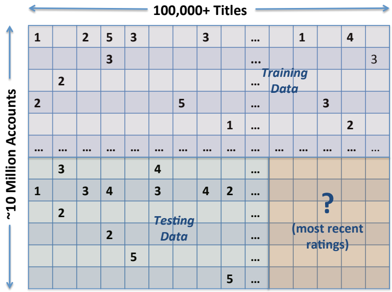
Mathematical Foundations

Data	Answers	Hidden Structure
X	Y	Z

Will NetFlix user 24601 like Transformers?

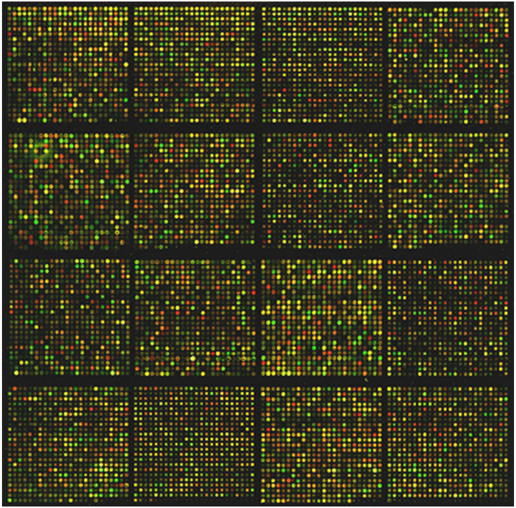
Will NetFlix user 24601 like Transformers?

How do you know?



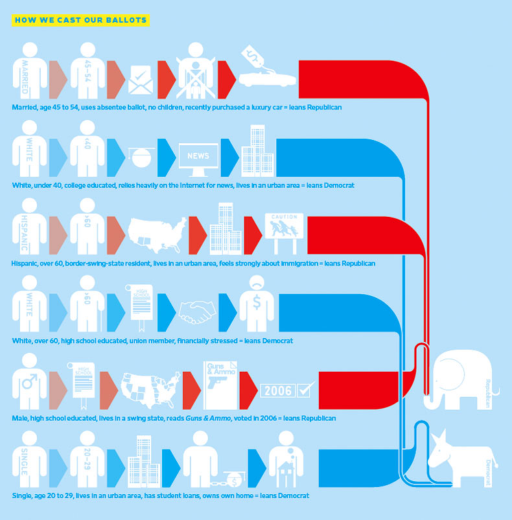
Group many images and determine the number of groups

Which genes are associated with a disease? How can expression values be used to predict survival?



Is it likely that this stock was traded based on illegal insider information?

Who will vote and for whom?



Is this spam?

Subject: CHARITY.
Date: February 4, 2008 10:22:25 AM EST
To: undisclosed-recipients:;
Reply-To: s.polla@yahoo.fr

Dear Beloved,
My name is Mrs. Susan Polla, from ITALY. If you are a christian and interested in charity please reply me at : (s.polla@yahoo.fr) for insight.
Respectfully,
Mrs Susan Polla.

Where are the faces?

Data contain patterns
that can help us solve problems.

This Course (Machine Learning)

We will study algorithms that find and exploit patterns in data.

- These algorithms draw on ideas from statistics and computer science.
- Applications include
  - natural science (e.g., genomics, neuroscience)
  - web technology (e.g., Google, NetFlix)
  - finance (e.g., stock prediction)
  - policy (e.g., predicting what intervention X will do)
  - and many others

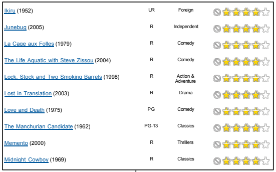
This Course (Machine Learning)

We will study algorithms that find and exploit patterns in data.

- Goal: fluency in thinking about modern machine learning problems.
- We will learn about a suite of tools in modern data analysis.
  - When to use them
  - The assumptions they make about data
  - Their capabilities, and their limitations
  - Theoretical guarantees
- We will learn a language and process for of solving data analysis problems. On completing the course, you will be able to learn about a new tool, apply it data, and understand the meaning of the result.

Basic idea behind everything we will study

1. **Collect or happen upon data.**
2. **Analyze it to find patterns.**
3. **Use those patterns to do something.**



Ikiru (1952)	UR	Foreign
Junebug (2005)	R	Independent
La Cage aux Folles (1979)	R	Comedy
The Life Aquatic with Steve Zissou (2004)	R	Comedy
Lock, Stock and Two Smoking Barrels (1998)	R	Action & Adventure
Lost in Translation (2003)	R	Drama
Love and Death (1975)	PG	Comedy
The Manchurian Candidate (1962)	PG-13	Classics
Memento (2000)	R	Thrillers
Midnight Cowboy (1969)	R	Classics

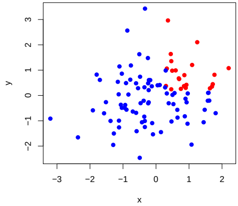
learning algorithm → predictor → 4.3 stars

How the ideas are organized

Of course, there is no one way to organize such a broad subject.
These concepts will recur through the course:

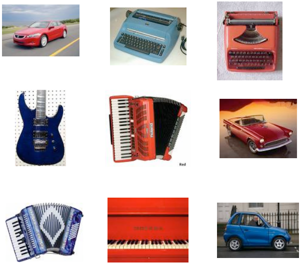
- Probabilistic foundations
- Supervised learning (more of this)
- Unsupervised learning (less of this)
- Methods that operate on discrete data (more of this)
- Methods that operate on continuous data (less of this)
- Representing data / feature engineering
- Evaluating models
- Understanding the assumptions behind the methods

Supervised vs. unsupervised methods



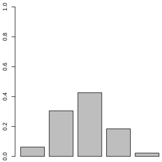
- **Supervised methods** find patterns in **fully observed** data and then try to predict something from **partially observed** data.
- For example, we might observe a collection of emails that are categorized into *spam* and *not spam*.
- After learning something about them, we want to take new email and automatically categorize it.

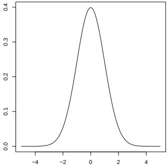
Supervised vs. unsupervised methods



- **Unsupervised methods** find **hidden structure** in data, structure that we can never formally observe.
- E.g., a museum has images of their collection that they want grouped by similarity into 15 groups.
- Unsupervised learning is more difficult to evaluate than supervised learning. But, these kinds of methods are widely used.

Discrete vs. continuous methods





- Discrete methods manipulate a finite set of objects
  - e.g., classification into one of 5 categories.
- Continuous methods manipulate continuous values
  - e.g., prediction of the change of a stock price.

One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Classification

SVM, naïve Bayes, logistic regression, boosting

One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Clustering

k-means, latent Dirichlet allocation

One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Regression

Linear Regression, Ridge Regression, Lasso

One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Dimensionality Reduction

...

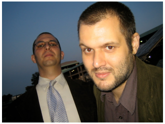
One useful grouping

	discrete	*continuous*
supervised	**classification**	**regression**
unsupervised	**clustering**	**dimensionality reduction**

Other

Reinforcement Learning, Ranking, Structured Prediction

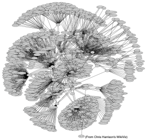
Data representation (feature engineering)



$\longrightarrow \langle 1.5, 3.2, -5.1, \dots, 4.2 \rangle$

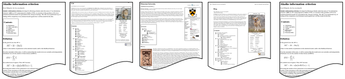
Republican nominee George Bush said he felt nervous as he voted today in his adopted home state of Texas, where he ended...

$\longrightarrow \langle 1, 0, 0, 0, 5, 0, 9, 3, 1, \dots, 0 \rangle$



$$\longrightarrow \begin{bmatrix} 1 & 0 & 1 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ 1 & 0 & 0 & \dots & 1 \\ & \dots & & & \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

Understanding assumptions



- The methods we'll study make **assumptions** about the data on which they are applied. E.g.,
  - Documents can be analyzed as a sequence of words;
  - or, as a "bag" of words.
  - Independent of each other;
  - or, as connected to each other
- What are the assumptions behind the methods?
- When/why are they appropriate?
- Much of this is an art

A Simple Example

- Suppose you're a big company monitoring the web
- Someone says something about your product (x)
- You want to know whether they're positive ($y = +1$) or negative ($y = -1$)

Train

Apple makes great laptops $\longrightarrow$ (+1)

Train

Apple makes great laptops → (+1)

Test

Apple makes great laptops

Train

Apple makes great laptops $\longrightarrow$ (+1)

Test

Apple really makes great laptops

Our (Usual) Assumption

- We have **training** examples $\{x_1, y_1\} \dots \{x_N, y_N\}$
- We have an unknown **test** example x without y
- What do we predict $h(x)$?

A simple solution

- Find something similar

A simple solution

- Find something similar

Discrete

$$d(x_1, x_2) = 1 - \frac{|x_1 \cap x_2|}{|x_1 \cup x_2|} \tag{1}$$

Continuous

$$d(x_1, x_2) = (\vec{x}_1 - \vec{x}_2)^2 \tag{2}$$

A simple solution

- Find something similar

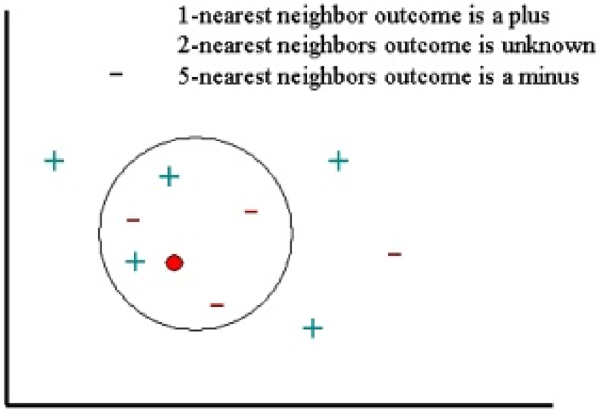
Discrete

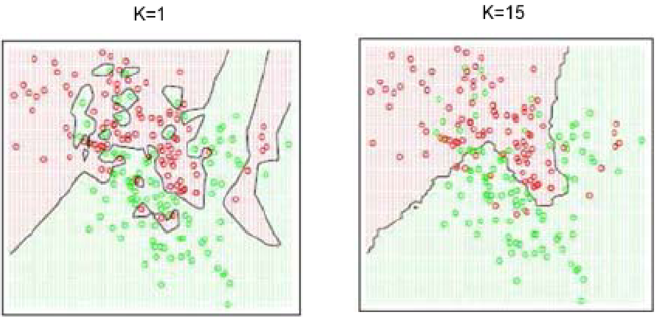
$$d(x_1, x_2) = 1 - \frac{|x_1 \cap x_2|}{|x_1 \cup x_2|} \tag{1}$$

Continuous

$$d(x_1, x_2) = (\vec{x}_1 - \vec{x}_2)^2 \tag{2}$$

- We can do better . . . look for the k closest and return the average y


1-nearest neighbor outcome is a plus
2-nearest neighbors outcome is unknown
5-nearest neighbors outcome is a minus


K=1
K=15

First Homework

- Implement *k*-nearest neighbors
- Acclimate you to the Python programming environment
- Introduce you to assignment submission

Next time . . .

- Probabilities
- Learning from data
  - Naïve Bayes
  - Logistic Regression