



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

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

Purchase histories

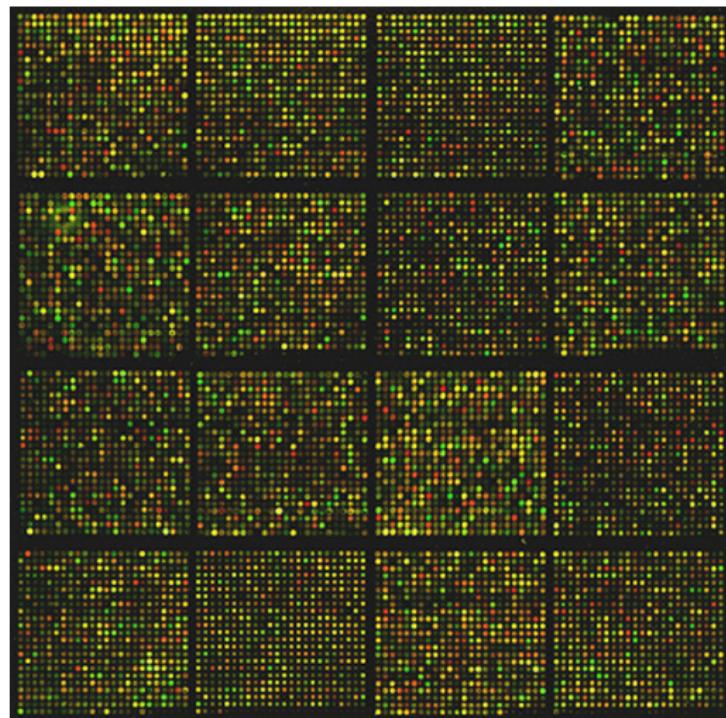
	Cheese			
0.5/0.51 lb	Cabot Vermont Cheddar	0.51 lb	\$7.99/lb	\$4.07
	Dairy			
1/1	Friendship Lowfat Cottage Cheese (16oz)		\$2.89/ea	\$2.89
1/1	Nature's Yoke Grade A Jumbo Brown Eggs (1 dozen)		\$1.49/ea	\$1.49
1/1	Santa Barbara Hot Salsa, Fresh (16oz)		\$2.69/ea	\$2.69
1/1	Stonyfield Farm Organic Lowfat 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



What can we do with data?

Genomics

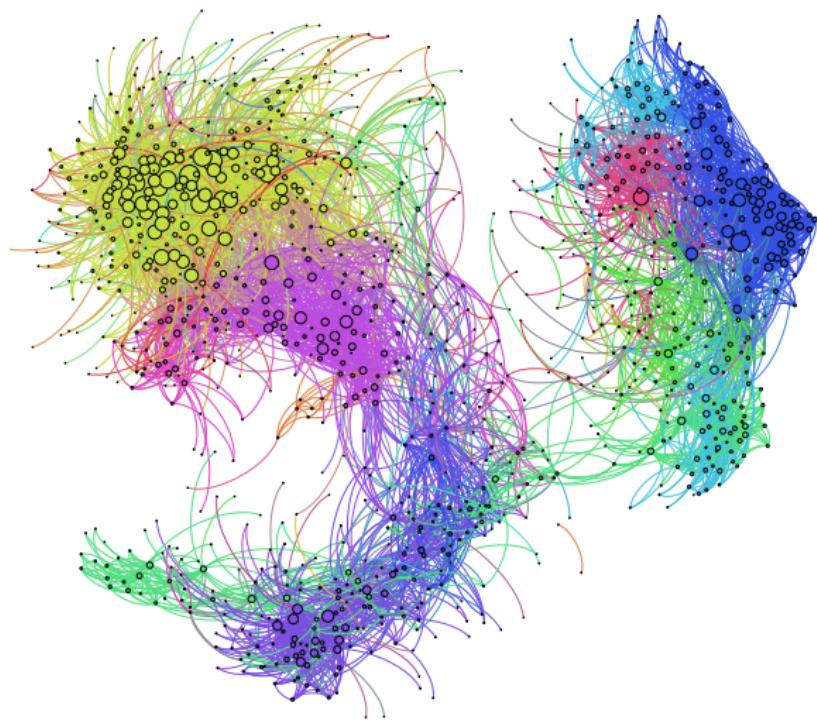


Neuroscience



What can we do with data?

Social networks



What can we do with data?

Finance



Data can help us solve problems.

Mathematical Foundations

Data

X

Answers

Y

Hidden Structure

Z

What can we do with data?

Will NetFlix user 24601 like Transformers?

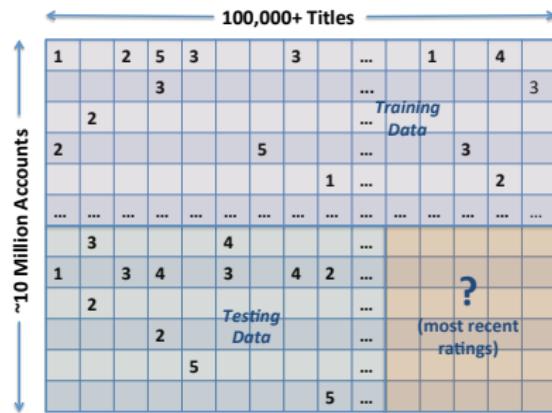


What can we do with data?

Will NetFlix user 24601 like Transformers?



How do you know?

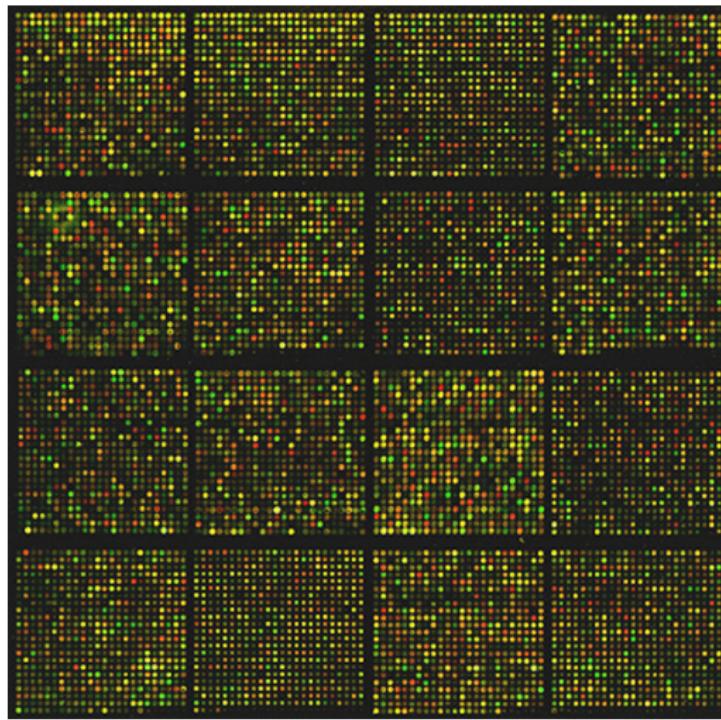


What can we do with data?

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?

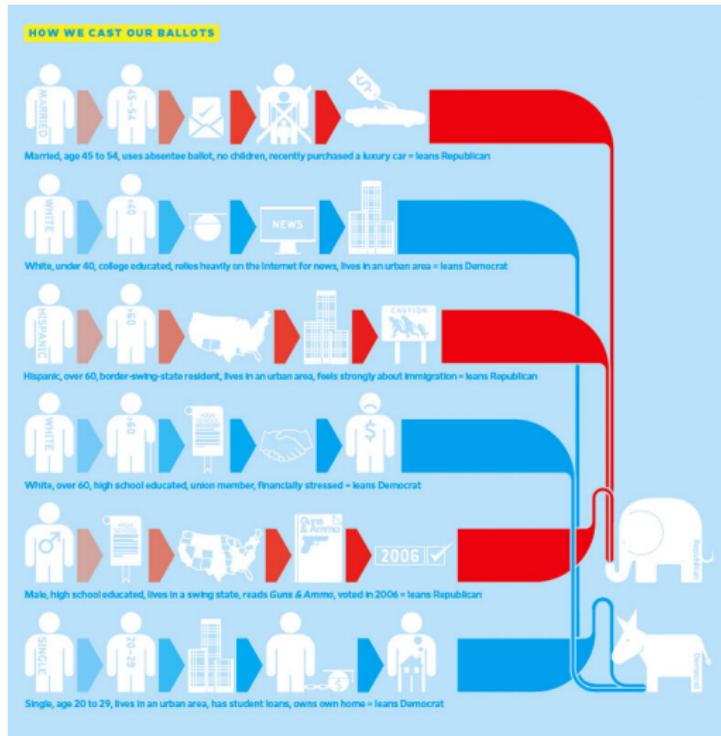


What can we do with data?

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.

What can we do with data?

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

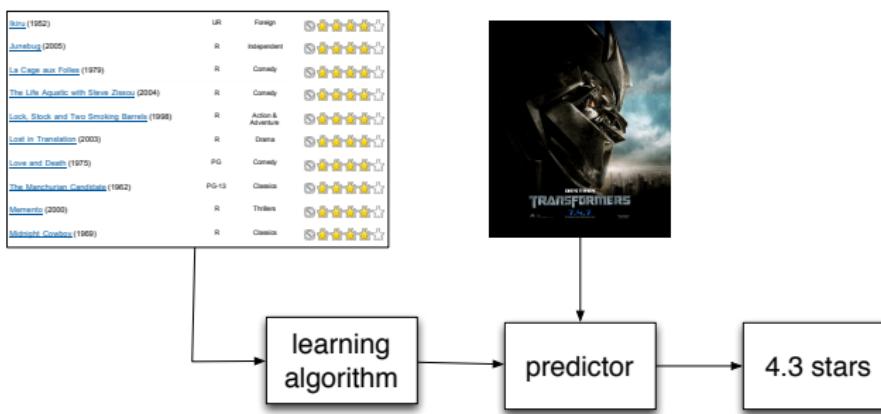
This Course (Machine Learning)

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 solving data analysis problems. On completing the course, you will be able to learn about a new tool, apply it to 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.

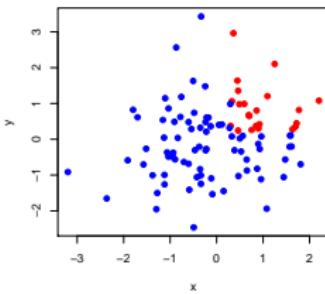


How the ideas are organized

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



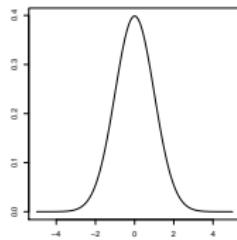
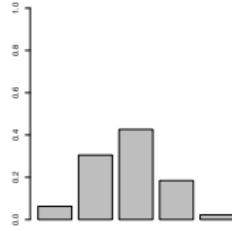
- **Supervised methods** find patterns in **fully observed** data and then try to predict something from **partially observed** data.
- For example, we might observe a collection of emails that are categorized into *spam* and *not spam*.
- After learning something about them, we want to take new email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

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

Classification

SVM, naïve Bayes, logistic regression, boosting

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

Clustering

k-means, latent Dirichlet allocation

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

Regression

Linear Regression, Ridge Regression, Lasso

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

Dimensionality Reduction

...

One useful grouping

	<i>discrete</i>	<i>continuous</i>
<i>supervised</i>	classification	regression
<i>unsupervised</i>	clustering	dimensionality reduction

Other

Reinforcement Learning, Ranking, Structured Prediction

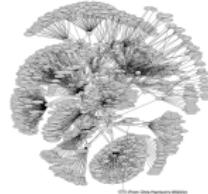
Data representation (feature engineering)



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

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

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



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

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 → (+1)

Train

Apple makes great laptops → (+1)

Test

Apple makes great laptops

Train

Apple makes great laptops → (+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

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

Continuous

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

A simple solution

- Find something similar

Discrete

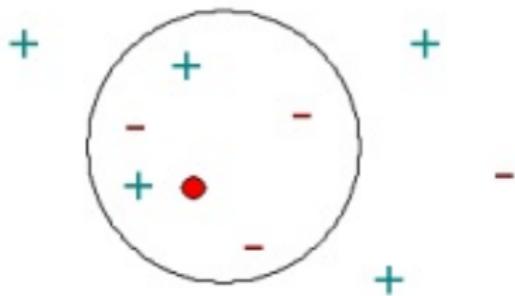
$$d(x_1, x_2) = 1 - \frac{|x_1 \cap x_2|}{|x_1 \cup x_2|} \quad (1)$$

Continuous

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

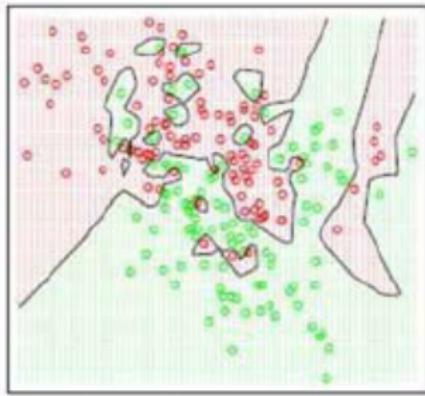
- We can do better . . . look for the k closest and return the average y

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

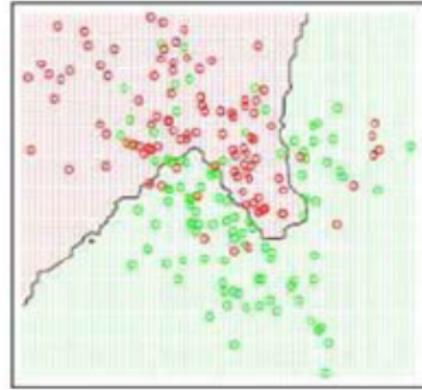


k -Nearest Neighbors

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