

Unsupervised & Semi-supervised Learning

John Blitzer, John DeNero, Dan Klein

1

Recap: Classification

- Classification systems:
 - Supervised learning
 - Make a prediction given evidence
 - We've seen several methods for this
 - Useful when you have labeled data



2

Clustering

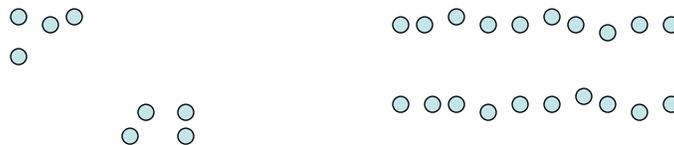
- Clustering systems:
 - Unsupervised learning
 - Detect patterns in unlabeled data
 - E.g. group emails or search results
 - E.g. find categories of customers
 - E.g. detect anomalous program executions
 - Useful when don't know what you're looking for
 - Requires data, but no labels
 - Often get gibberish



3

Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



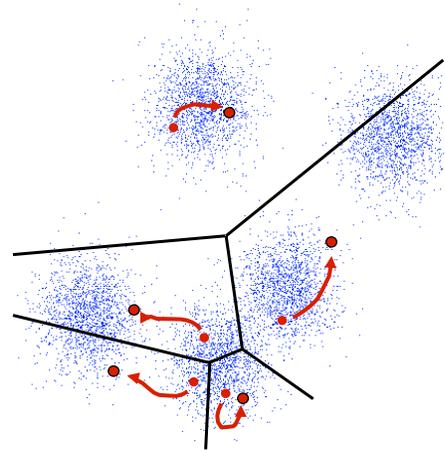
- What could “similar” mean?
 - One option: small (squared) Euclidean distance

$$\text{dist}(x, y) = (x - y)^T (x - y) = \sum_i (x_i - y_i)^2$$

4

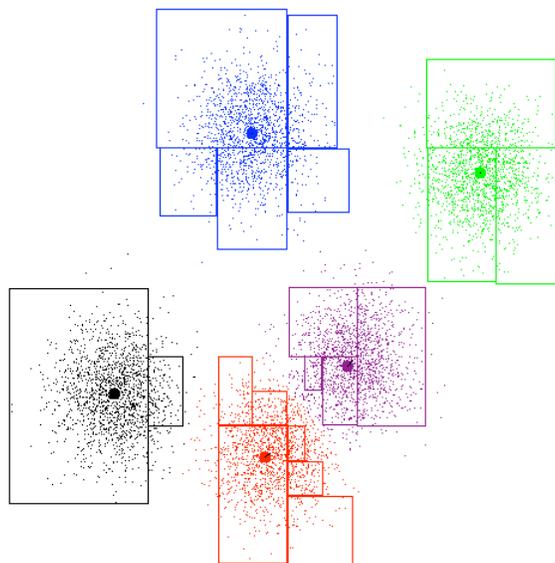
K-Means

- An iterative clustering algorithm
 - Pick K random points as cluster centers (means)
 - Alternate:
 - Assign data instances to closest mean
 - Assign each mean to the average of its assigned points
 - Stop when no points' assignments change



5

K-Means Example



6

Example: K-Means

- [web demo]
 - <http://www.cs.washington.edu/research/imagedatabase/demo/kmcluster/>

7

K-Means as Optimization

- Consider the total distance to the means:

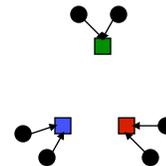
$$\phi(\{x_i\}, \{a_i\}, \{c_k\}) = \sum_i \text{dist}(x_i, c_{a_i})$$

points assignments means

- Each iteration reduces phi

- Two stages each iteration:

- Update assignments: fix means c , change assignments a
- Update means: fix assignments a , change means c



8

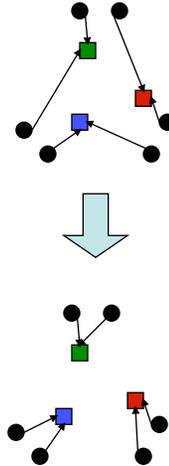
Phase I: Update Assignments

- For each point, re-assign to closest mean:

$$a_i = \operatorname{argmin}_k \operatorname{dist}(x_i, c_k)$$

- Can only decrease total distance phi!

$$\phi(\{x_i\}, \{a_i\}, \{c_k\}) = \sum_i \operatorname{dist}(x_i, c_{a_i})$$



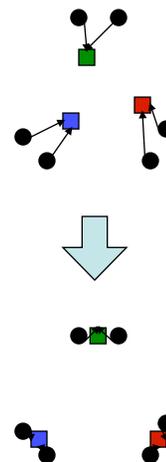
9

Phase II: Update Means

- Move each mean to the average of its assigned points:

$$c_k = \frac{1}{|\{i : a_i = k\}|} \sum_{i: a_i = k} x_i$$

- Also can only decrease total distance... (Why?)
- Fun fact: the point y with minimum squared Euclidean distance to a set of points $\{x\}$ is their mean

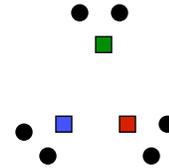


10

Initialization

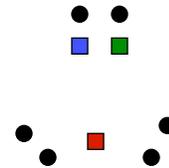
- K-means is non-deterministic

- Requires initial means
- It does matter what you pick!



- What can go wrong?

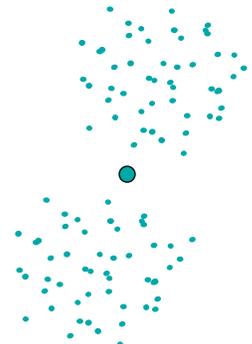
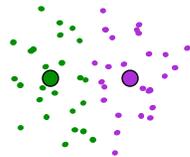
- Various schemes for preventing this kind of thing: variance-based split / merge, initialization heuristics



11

K-Means Getting Stuck

- A local optimum:



Why doesn't this work out like the earlier example, with the purple taking over half the blue?

12

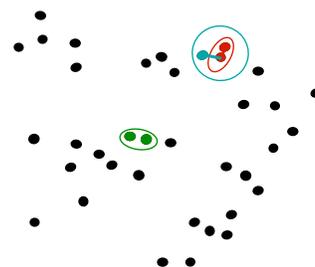
K-Means Questions

- Will K-means converge?
 - To a global optimum?
- Will it always find the true patterns in the data?
 - If the patterns are very very clear?
- Will it find something interesting?
- Do people ever use it?
- How many clusters to pick?

13

Agglomerative Clustering

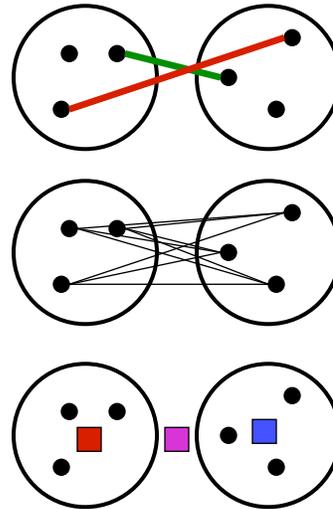
- **Agglomerative clustering:**
 - First merge very similar instances
 - Incrementally build larger clusters out of smaller clusters
- **Algorithm:**
 - Maintain a set of clusters
 - Initially, each instance in its own cluster
 - Repeat:
 - Pick the two **closest** clusters
 - Merge them into a new cluster
 - Stop when there's only one cluster left
- Produces not one clustering, but a family of clusterings represented by a **dendrogram**



14

Agglomerative Clustering

- How should we define “closest” for clusters with multiple elements?
- Many options
 - Closest pair (single-link clustering)
 - Farthest pair (complete-link clustering)
 - Average of all pairs
 - Ward’s method (min variance, like k-means)
- Different choices create different clustering behaviors



15

Clustering Application

Google News Search and browse 25,000 news sources updated continuously.

World | **U.S.**

Heavy Fighting Continues As Pakistan Army Battles Taliban
 Voice of America - 10 hours ago
 By Barry Newhouse Pakistan's military said its forces have killed 55 to 60 Taliban militants in the last 24 hours in heavy fighting in Taliban-held areas of the northwest. [Pakistan troops battle Taliban militants for fourth day](#) guardian.co.uk
[Army: 55 militants killed in Pakistan fighting](#) The Associated Press
[Christian Science Monitor](#) - [CNN International](#) - [Bloomberg](#) - [New York Times](#)
[all 3,824 news articles »](#)

Sri Lanka admits bombing safe haven
 guardian.co.uk - 3 hours ago
 Sri Lanka has admitted bombing a "safe haven" created for up to 150,000 civilians fleeing fighting between Tamil Tiger fighters and the army. [Chinese billions in Sri Lanka fund battle against Tamil Tigers](#) Times Online
[Huge Humanitarian Operation Under Way in Sri Lanka](#) Voice of America
[BBC News](#) - [Reuters](#) - [AFP](#) - [Xinhua](#)
[all 2,492 news articles »](#)

Business

Buffett Calls Investment Candidates' 2008 Performance Subpar
 Bloomberg - 2 hours ago
 By Hugh Son, Erik Holm and Andrew Frye May 2 (Bloomberg) -- Billionaire Warren Buffett said all of the candidates to replace him as chief investment officer of Berkshire Hathaway Inc. failed to beat the 38 percent decline of the Standard & Poor's 500 ... [Buffett offers bleak outlook for US newspapers](#) Reuters
[Buffett: Unit CEO say through embarrassment](#) MarketWatch
[CNBC](#) - [The Associated Press](#) - [guardian.co.uk](#)
[all 1,454 news articles »](#) [BRK-A](#)

Chrysler's Fall May Help Administration Reshape GM
 New York Times - 5 hours ago
 Auto task force members, from left, Treasury's Ron Bloom and Gene Sperling, Labor's Edward Montgomery, and Steve Rattner. BY DAVID E. GANGER and BILL VLASIC WASHINGTON - Fresh from pushing Chrysler into bankruptcy, President Obama and his economic team ... [Comment by Gary Chaison](#) Prof of Industrial Relations, Clark University
[Bankruptcy reality sets in for Chrysler workers](#) Detroit Free Press
[Washington Post](#) - [Bloomberg](#) - [CNNMoney.com](#)
[all 11,028 news articles »](#) [OTC-FIAT](#) - [BIT-PR](#) - [GM](#)

Weekend Opinionator: Souter, Specter and the Future of the GOP
 New York Times - 48 minutes ago
 By Tobin Harshaw An odd week. While Barack Obama celebrated his 100th day in office the headlines were pretty much dominated by the opposition party, albeit not in the way many Republicans would have liked. [US Supreme Court Vacancy An Early Test For Sen Specter](#) Wall Street Journal
[Letters: Arlen Specter, Notre Dame, Chrysler](#) Houston Chronicle
[The Associated Press](#) - [Kansas City Star](#) - [Philadelphia Inquirer](#) - [Bangor Daily News](#)
[all 401 news articles »](#)

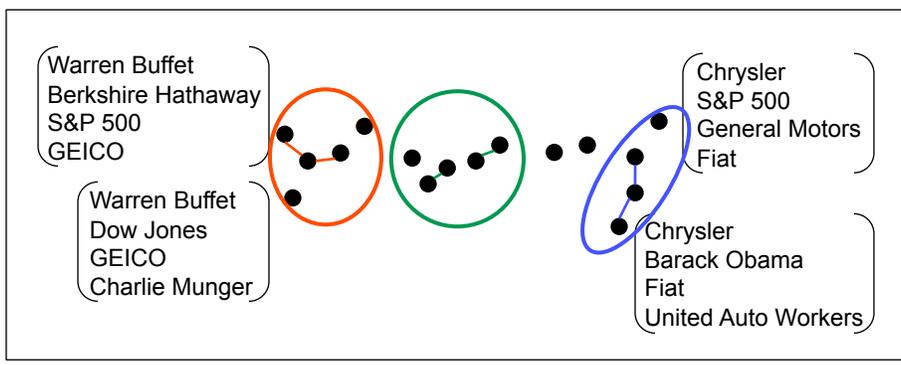
Joe Biden, the Flu and You
 New York Times - 48 minutes ago
 By GAIL COLLINS The swine flu scare has made it clear why Barack Obama picked Joe Biden for vice president. David Brooks and Gail Collins talk between columns. [After his flu warning, Biden takes the train home](#) The Associated Press
[Biden to visit Balkan states in mid-May](#) Washington Post
[AFP](#) - [Christian Science Monitor](#) - [Bizjournals.com](#) - [Voice of America](#)
[all 1,506 news articles »](#)

Top-level categories: supervised classification

Story groupings: unsupervised clustering

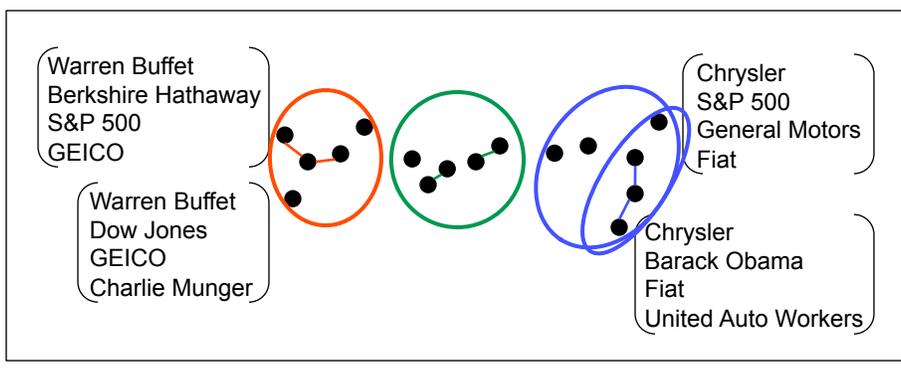
Step 1: Agglomerative Clustering

- Separate clusterings for each global category
- Represent documents as vectors
 - Millions of dimensions (1 for each proper noun)
- How do we know when to stop?



Step 2: K-means clustering

- Initialize means to centers from agglomerative step
- Why might this be a good idea?
 - Guaranteed to decrease squared-distance from cluster means
 - Helps to “clean up” points that may not be assigned appropriately



Semi-supervised learning

- For a particular task, labeled data is always better than unlabeled
 - Get a correction for every mistake

- But labeled data is usually much more expensive to obtain
 - Google News: Manually label news story clusters every 15 minutes?
 - Other examples? Exceptions?

- Combine labeled and unlabeled data to build better models



Sentiment Analysis

Elias Corner
 2402 31st St
 Long Island City, NY 11102
 (718) 932-1510
[nymag.com](#)

Get Directions: [To here](#) - [From here](#)
[Add or edit your business](#)

Overview Details (8) Reviews (24) Photos (3)

topic summaries

general comments ★★★★★ it was awful...Real typical Greece
 fish ★★★★★ fresh fish...Very simple grilled fish
 food ★★★★★ good food... Nice fresh salads...
 service ★★★★★ worst service...casual service...
 decor ★★★★★
 value ★★★★★

topic details

Live Search digital camera

Products
 See also: [Web](#), [Images](#), [Videos](#), [News](#), [Maps](#), [More](#)

Canon PowerShot SD1000 Digital ELPH - digital camera, 7.1MP, 3x ...

\$129 - \$186 Compare prices (3) 2% - 9% cashback

★★★★★ User reviews (756)
 ★★★★★ Expert reviews (2)

Canon looked to the very first ELPH for inspiration when designing the PowerShot SD1000 Digital ELPH, and came up with a quintessential iteration of the icon: slim, clean-lined and fully flat. Inside, the SD1000 Digital ELPH looks... [More...](#)

User reviews | Product details | Expert reviews | Compare prices

All user reviews

All user reviews
 View by: All | Highest rating | Lowest rating

Size (325 comments)
 97% positive

Ease Of Use (292 comments)
 95% positive

Photo Quality (247 comments)
 94% positive

Affordability (112 comments)
 91% positive

Speed (97 comments)
 88% positive

wonderful camera
 I think this is an excellent product. The only thing I don't like is I'm not able to edit the picture after it has been taken on the camera itself. It takes excellent quality...
 ★★★★★ shehas109 www.ebay.com 3/26/2009

GREAT SELLER!
 First of all, I had my camera stolen along with my phone a week ago. I already new that I loved my camera and I wanted the exact same one. I found this one on ebay. It was a great...
 ★★★★★ birdgolfer88 www.ebay.com 3/21/2009

READ DESCRIPTION CAREFULLY - ITEMS NOT NEW
 THE SELLER DOES NOT FULLY REVEAL THAT HIS

Other companies: <http://www.jodange.com> , <http://www.brandtology.com> , ...

Sentiment Classification

Product Review

Linear classifier (perceptron)

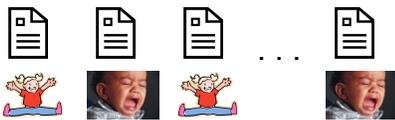


Positive



Negative

Supervised learning problem



Running with Scissors: A Memoir

Title: Horrible book, horrible.

This book was horrible. I read half of it, suffering from a headache the entire time, and eventually i lit it on fire. One less copy in the world...don't waste your money. I wish i had the time spent reading this book back so i could use it for better purposes. This book wasted my life

Features for Sentiment Classification

This book was **horrible**. I read half of it, **suffering** from a **headache** the entire time, and eventually i lit it **on fire**. One less copy in the world...**don't waste** your money. I wish i had the time spent reading this book back so i could use it for better purposes. This book **wasted** my life

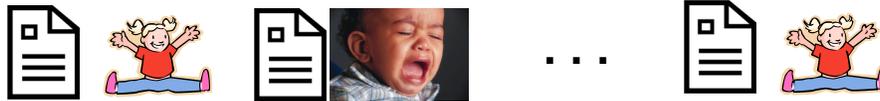
Recall perceptron classification rule:

$$y = \arg \max_y w_y \cdot x$$

Features: counts of particular words & bigrams

Domain Adaptation

Training data: **labeled book reviews**



Test data: **unlabeled kitchen appliance reviews**



Semi-supervised problem: Can we build a good classifier for kitchen appliances?

Books & Kitchen Appliances

Running with Scissors: A Memoir

Title: Horrible book, horrible.

This book was horrible. I **read half**

of it. **suffering from a headache** the

Arata Dopp Feyer, Gihoma &

Blakk

Title: It does not work well...

I don't know what the **fallen** fire

Error increase: 13% → 26%

fire. One less copy in the

world...don't waste your money. I wish i had the time spent reading this book back so i could use it for better purposes. This book wasted my life

senyrd condone **due to a defective**

closure. The **flame** close initially, but it **blows** away a few **long** it stays longer. **It's old** **pl** **will not** get this **open** **is**ing this one again.

Handling Unseen Words

- (1) **Unsupervised**: Cluster words based on context
- (2) **Supervised**: Use clusters in place of the words themselves

Clustering intuition: Contexts for the word **defective**

Unlabeled **kitchen** contexts

Unlabeled **books** contexts

- Do **not buy** the Shark portable steamer Trigger mechanism is **defective**.
- the very nice lady assured me that I must have a **defective** set What a **disappointment!**

- The book is so **repetitive** that I found myself yelling I will definitely **not buy** another.
- A **disappointment** Ender was talked about for **<#>** **pages** altogether.

Clustering: Feature Vectors for Words

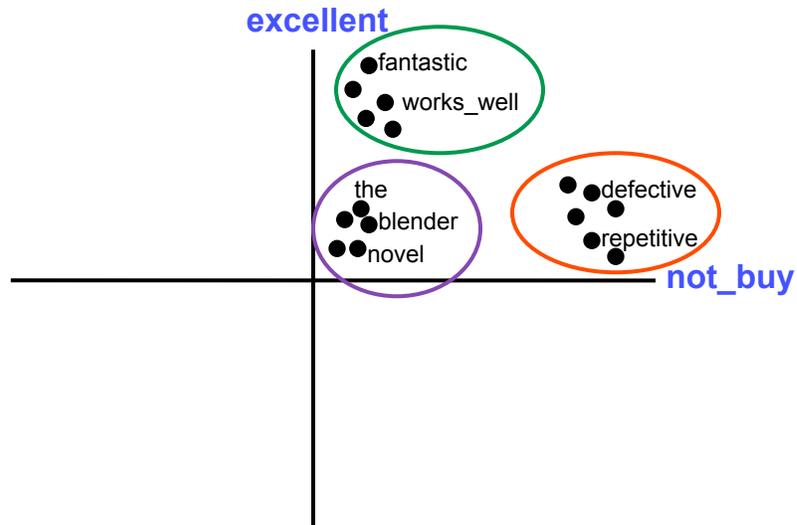
Approximately 1000 pivots, 1 million feature words

Feature Words

		fascinating		defective		repetitive
Pivots	excellent	1	...	0	...	0

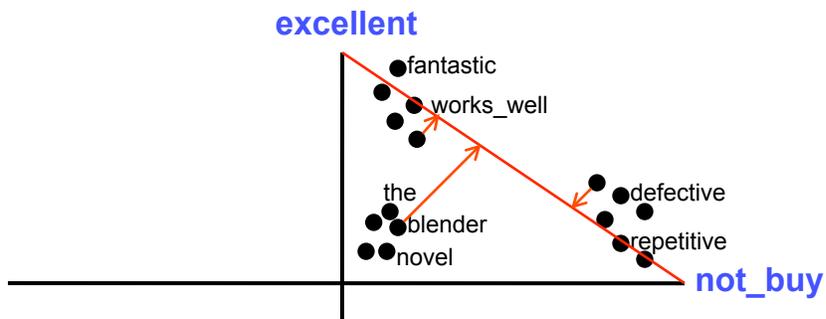
	not buy	0	...	2	...	1
	awful	0		2		0
	terrible	0	...	1	...	3

K-Means in Pivot Space



27

Real-valued Linear Projections



Position along the line gives a **real-valued** soft notion of “polarity” for each word

28

