Unsupervised & Semi-supervised Learning

John Blitzer, John DeNero, Dan Klein

Recap: Classification

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

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

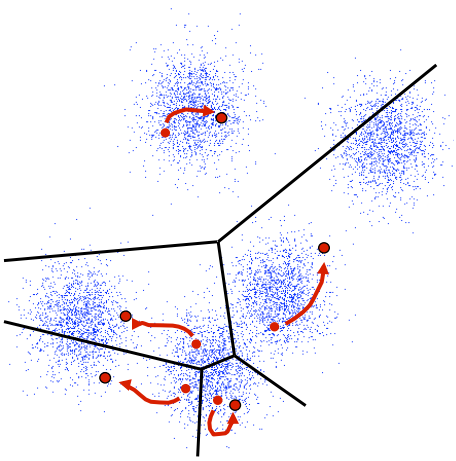
- What could "similar" mean?
  - One option: small (squared) Euclidean distance

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

4

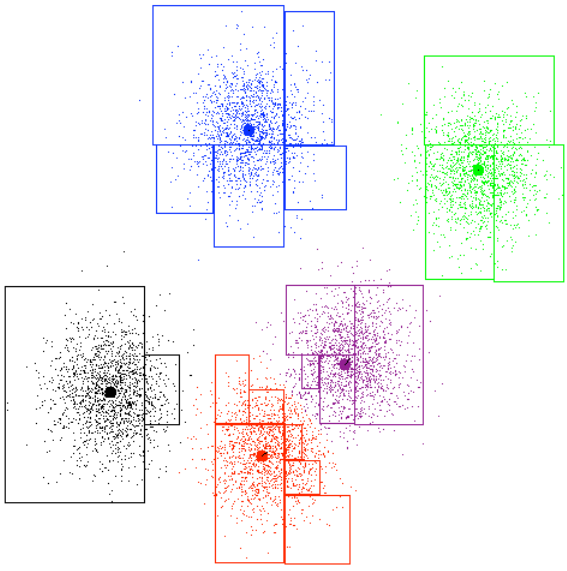
K-Means

- 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/

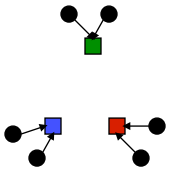
K-Means as Optimization

- 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

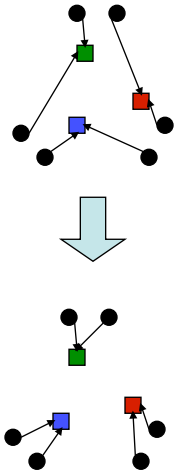
Phase I: Update Assignments

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

$$a_i = \underset{k}{\text{argmin}}\, \text{dist}(x_i, c_k)$$

- Can only decrease total distance phi!

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



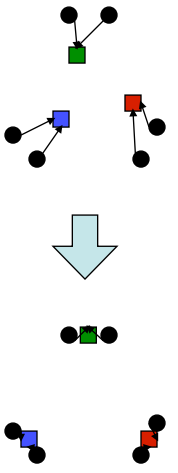
9

Phase II: Update Means

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

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

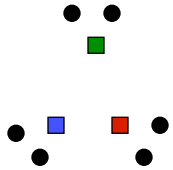
- Also can only decrease total distance… (Why?)
- Fun fact: the point y with minimum squared Euclidean distance to a set of points {x} is their mean

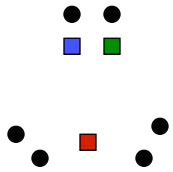


10

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

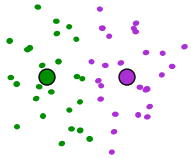


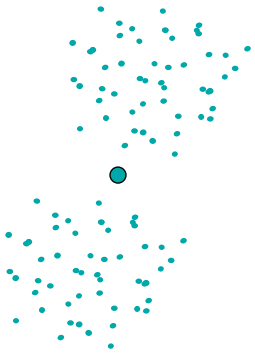


11

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

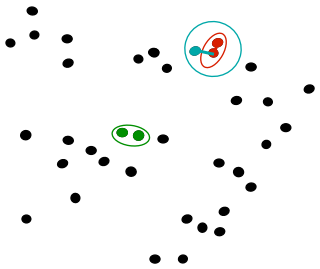
K-Means Questions

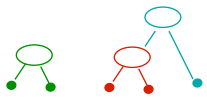
- Will K-means converge?
  - To a global optimum?
- Will it always find the true patterns in the data?
  - If the patterns are very very clear?
- Will it find something interesting?
- Do people ever use it?
- How many clusters to pick?

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

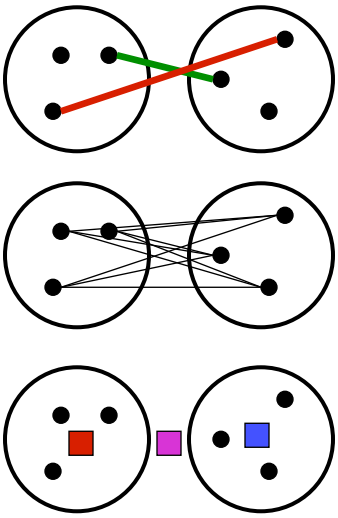




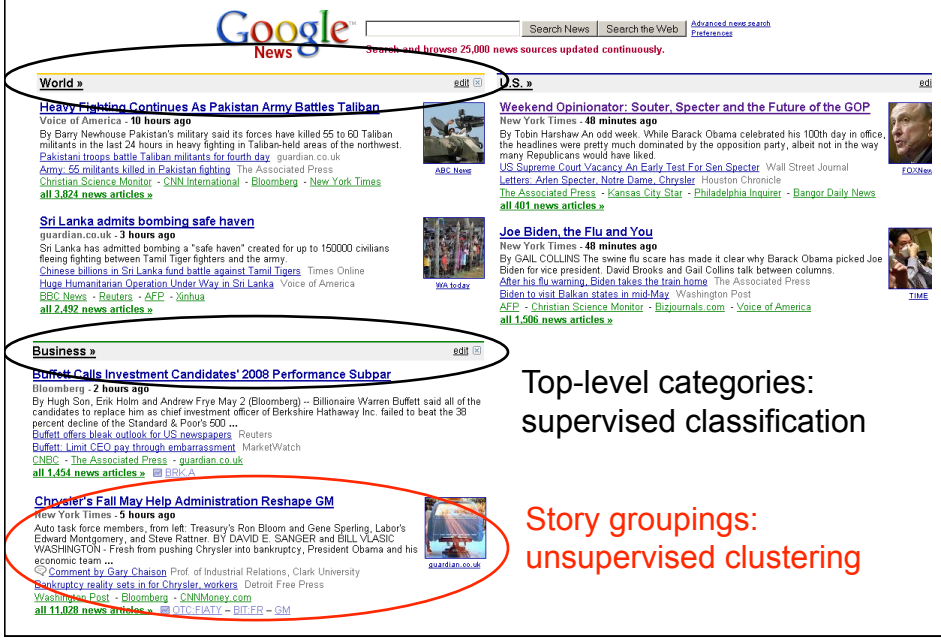
14

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

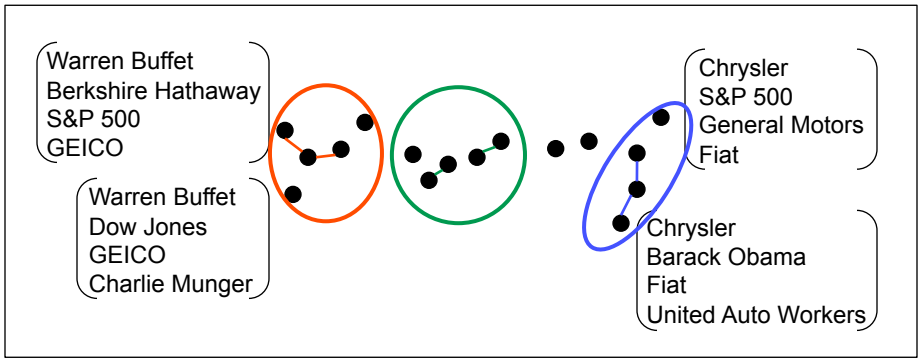


Clustering Application



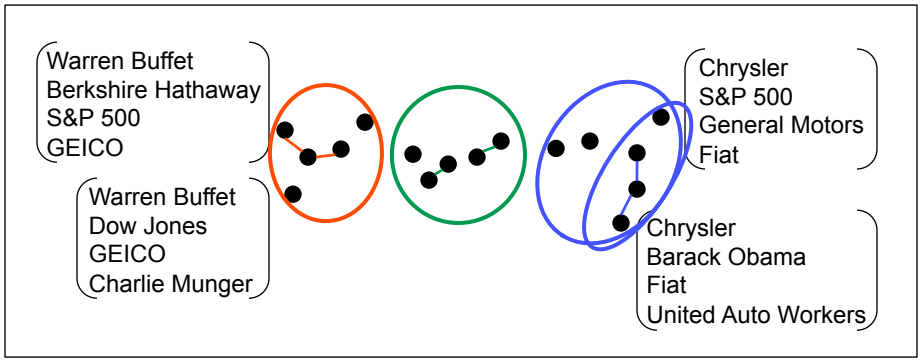
Step 1: Agglomerative Clustering

- 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?



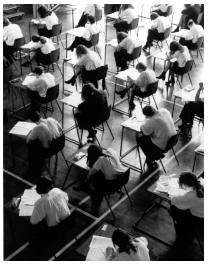
Step 2: K-means clustering

- 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



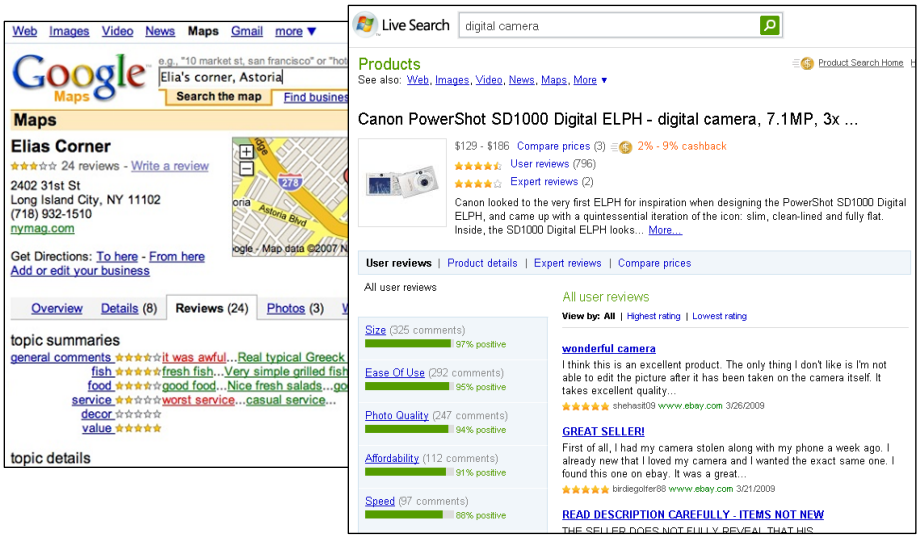
Semi-supervised learning

- **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



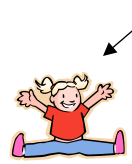
Other companies: http://www.jodange.com , http://www.brandtology.com , …

20

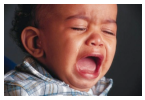
Sentiment Classification

Product Review

Linear classifier (perceptron)

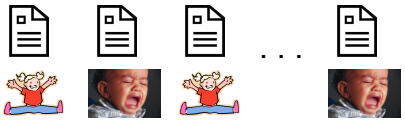


Positive



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:

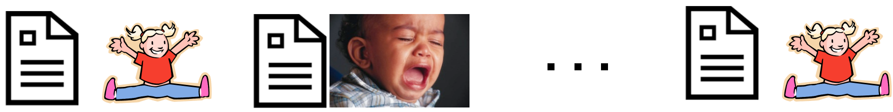
$$y = \arg\max_{y} \ \mathbf{w}_y \cdot \mathbf{x}$$

Features: counts of particular words & bigrams

22

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	**Avante Deep Fryer, Chrome & Black** **Title: lid does not work well...** I love the way the Tefal deep fryer

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	second one due to a defective lid closure. The lid may close initially, but after a few uses it no longer stays closed. I will not be purchasing 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 book is so **repetitive** that I found myself yelling …. I will definitely **not buy** another.
• the very nice lady assured me that I must have a **defective** set …. What a **disappointment**!	• A **disappointment** …. Ender was talked about for **<#> pages** altogether.

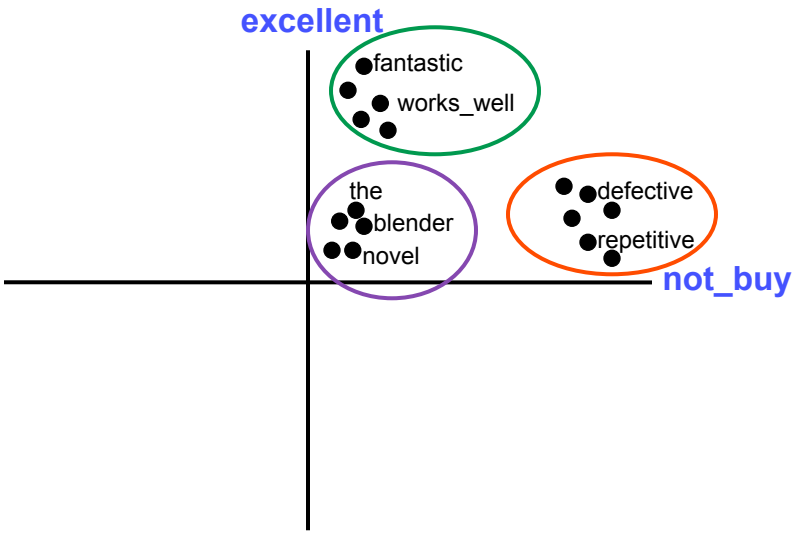
Clustering: Feature Vectors for Words

Approximately 1000 pivots, 1 million feature words

Feature Words

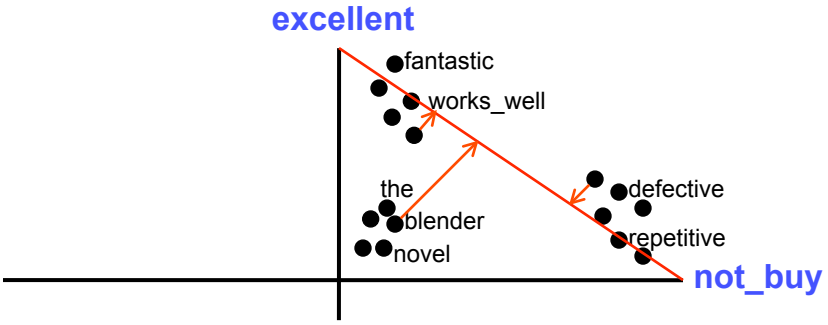
Pivots	fascinating	…	defective	…	repetitive
excellent	1	…	0	…	0
…	…		…		…
not buy	0	…	2	…	1
awful	0		2		0
terrible	0	…	1	…	3

K-Means in Pivot Space



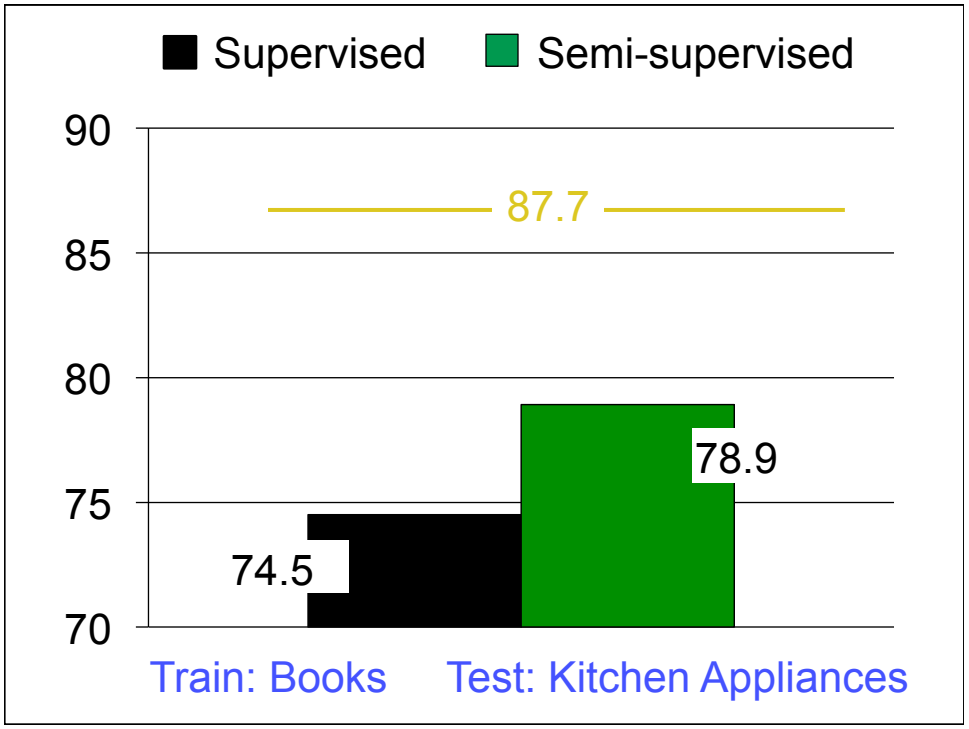
27

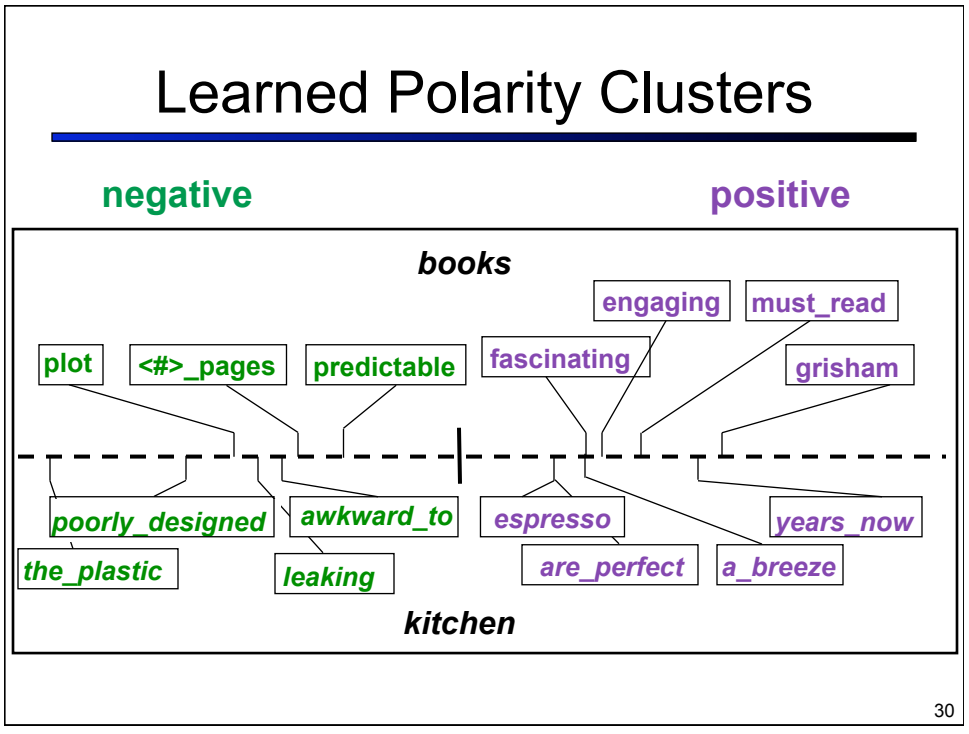
Real-valued Linear Projections



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

28


Supervised
Semi-supervised
90
85
80
75
70
87.7
78.9
74.5
Train: Books
Test: Kitchen Appliances


Learned Polarity Clusters
negative
positive
books
engaging
must_read
plot
<#>_pages
predictable
fascinating
grisham
poorly_designed
awkward_to
espresso
years_now
the_plastic
leaking
are_perfect
a_breeze
kitchen
30