# CSE 446: Clustering and EM
Winter 2012

Daniel Weld

Slides adapted from Carlos Guestrin, Dan Klein & Luke Zettlemoyer

## Machine Learning

- Supervised Learning
  - Parametric
    - Y Continuous
      - Gaussians
        Learned in closed form
      - Linear Functions
        1. Learned in closed form
        2. Using gradient descent
    - Y Discrete
      - Decision Trees
        Greedy search; pruning
      - Probability of class | features
        1. Learn P(Y), P(X|Y); apply Bayes
        2. Learn P(Y|X) w/ gradient descent
  - Non-parametric
    - Non-probabilistic
      - Linear: perceptron gradient descent
      - Nonlinear: neural net: backprop
- Unsupervised Learning
- Reinforcement Learning

2

## Machine Learning

- Supervised Learning
  - Parametric
  - Non-parametric
- Unsupervised Learning
- Reinforcement Learning

3

## Outline

| | |
|---|---|
| Fri | K-means & Agglomerative Clustering |
| Mon | Expectation Maximization (EM) |
| Wed | Principle Component Analysis (PCA) |
| Fri | Markov Decision Processes (MDPs) |
| Mon | Reinforcement Learning (RL) |
| Wed | Instance-Based Learning |
| Fri | SVMs & Summary |

4

## Overview of Learning

Type of Supervision
(eg, Experience, Feedback)

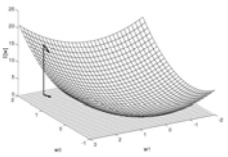
What is Being Learned?

| | Labeled Examples | Reward | Nothing |
|---|---|---|---|
| Discrete Function | Classification | | Clustering PCA |
| Continuous Function | Regression | | |
| Policy | Apprenticeship Learning | Reinforcement Learning | |

5

## Key Perspective on Learning

- Learning as Optimization
  - Closed form
  - Greedy search
  - Gradient ascent
- Loss Function
  - Error + regularization



6

## Clustering

Clustering systems:

- Unsupervised learning
- Requires data, but no labels
- Detect patterns eg in
  - Group emails or search results
  - Customer shopping patterns
  - Program executions (intrusion detection)
- Useful when don't know what you're looking for
- But: often get gibberish

## Clustering

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

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

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

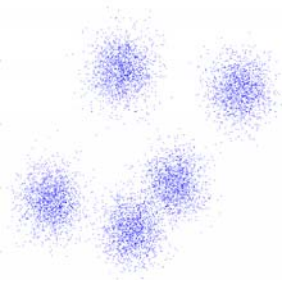
## Clustering Methods

- K-means
- Agglomerative clustering
- EM

9

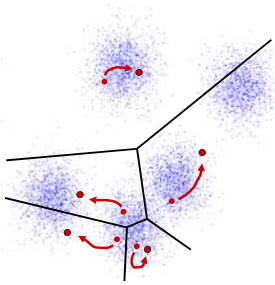
## K-Means

- 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

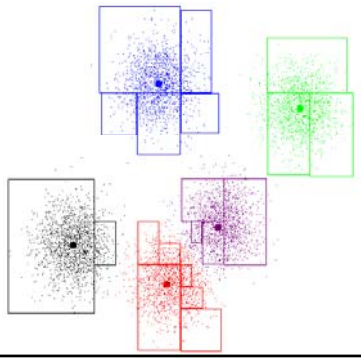


## K-Means

- 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



## K-Means Example



## Example: K-Means for Segmentation

K=2

Original

## Example: K-Means for Segmentation

K=2

K=3

Original

## Example: K-Means for Segmentation

K=2

K=3

K=10

Original

4%

8%

17%

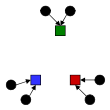
## K-Means as Optimization

- 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

- Two stages each iteration:
  - Update assignments: fix means c, change assignments a
  - Update means: fix assignments a, change means c



- *Coordinate gradient ascent* on Φ
- Will it converge?
  - Yes!, if you can argue that each update can't increase Φ

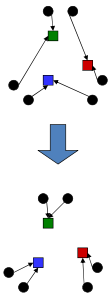
## Phase I: Update Assignments

- 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!

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

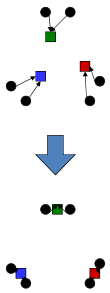


## Phase II: Update Means

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

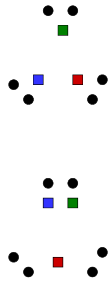
$$c_k = \frac{1}{|\{i : a_i = k\}|} \sum_{i:a_i=k} x_i$$

- 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



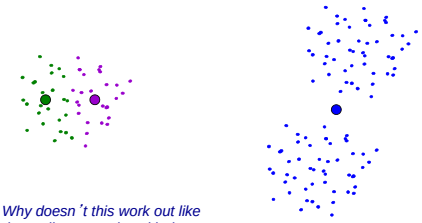
## Initialization

- 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



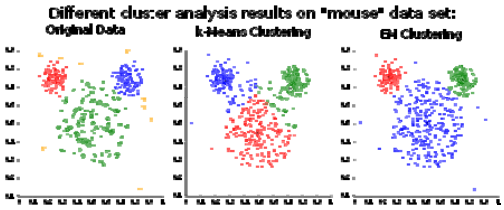
## K-Means Getting Stuck

A local optimum:



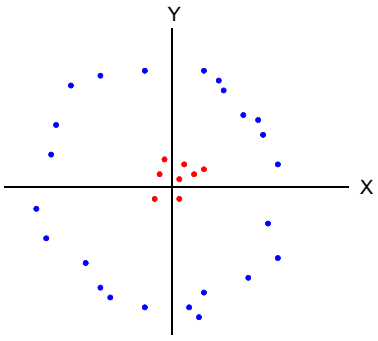
*Why doesn´t this work out like the earlier example, with the purple taking over half the blue?*

## Preference for Equally Sized Clusters



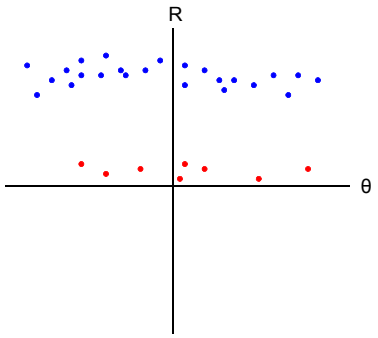



## Another Example






## Another Example



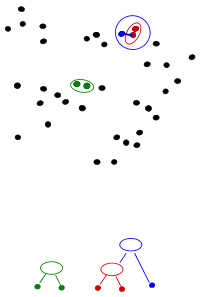



## K-Means Questions

- 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?
- Runtime?
- Do people ever use it?
- How many clusters to pick?

## Agglomerative Clustering

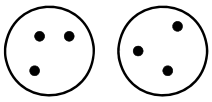
- Agglomerative clustering:
  - First merge very similar instances
  - Incrementally build larger clusters out of smaller clusters
- Algorithm:
  - Maintain a set of clusters
  - Initially, each instance in its own cluster
  - Repeat:
    - Pick the two closest clusters
    - Merge them into a new cluster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





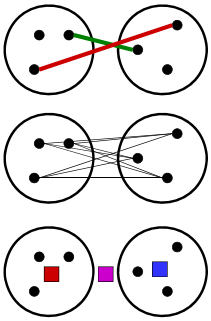
## Agglomerative Clustering

- How should we define "closest" for clusters with multiple elements?

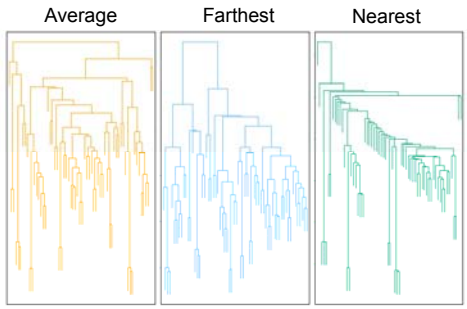


## Agglomerative Clustering

- 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 Behavior




Mouse tumor data from [Hastie]

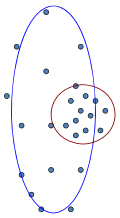
28

## Agglomerative Clustering Questions

- Will agglomerative clustering converge?
  - To a global optimum?

## Agglomerative Clustering Questions

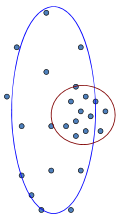
- Will agglomerative clustering converge?
  - To a global optimum?



## Agglomerative Clustering Questions

- Will agglomerative clustering converge?
  - To a global optimum?
- Will it always find the true patterns in the data?
- Do people ever use it?
- How many clusters to pick?

## Reconsidering "**hard assignments**"?



- Clusters may overlap
- Some clusters may be "wider" than others
- Distances can be deceiving!

## Acknowledgements


- K-means & Gaussian mixture models presentation contains material from excellent tutorial by Andrew Moore:
  - http://www.autonlab.org/tutorials/
- K-means Applet:
  - http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletKM.html
- Gaussian mixture models Applet:
  - http://www.neurosci.aist.go.jp/%7Eakaho/MixtureEM.html