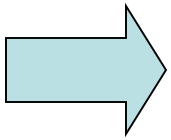
Unsupervised learning, Clustering

CS434

Unsupervised learning and pattern discovery

So far, our data has been in this form:

$x_1^1, x_2^1, x_3^1, \ldots, x_m^1$	y^1
$x_1^2, x_2^2, x_3^2, \ldots, x_m^2$	y^2
…	
…	
$x_1^n, x_2^n, x_3^n, \ldots, x_m^n$	y^n



We will be looking at ***unlabeled data:***

$x_1^1, x_2^1, x_3^1, \ldots, x_m^1$
$x_1^2, x_2^2, x_3^2, \ldots, x_m^2$
…
…
$x_1^n, x_2^n, x_3^n, \ldots, x_m^n$

What do we expect to learn from such data?

I have tons of data (web documents, gene data, etc) and need to:

organize it better – e.g., find subgroups

understand it better – e.g., understand interrelationships

find regular trends in it – e.g., If A and B, then C

The web organized by topic into categories.

Arts
Movies, Music, Television, ...

Business
Companies, Finance, Jobs, ...

Computers
Internet, Hardware, Software, ...

Games
Board, Roleplaying, Video, ...

Health
Alternative, Fitness, Medicine, ...

Home
Consumers, Homeowners, Family, ...

Kids and Teens
Computers, Entertainment, School, ...

News
Media, Newspapers, Current Events, ...

Recreation
Food, Outdoors, Travel, ...

Reference
Education, Libraries, Maps, ...

Regional
Asia, Europe, North America, ...

Science
Biology, Psychology, Physics, ...

Shopping
Autos, Clothing, Gifts, ...

Society
Issues, People, Religion, ...

Sports
Basketball, Football, Soccer, ...

World
Deutsch, Español, Français, Italiano, Japanese, Korean, Nederlands, Polska, Svenska, ...

Music

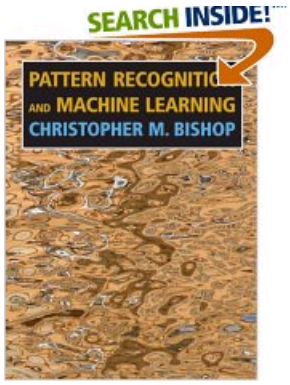
Arts > Music

Categories

Anime (61)
Anti-Music (172)
Arranging (2)
Awards (52)
Bands and Artists (47326)
Business (5322)
Charts (109)
Chats and Forums (242)
Classifieds (37)
Clubs and Venues (435)
Collecting (220)
Comedians (128)
Composition (16836)
Computers (38)
Concerts and Events (507)
Directories (120)
Disabled (17)
DJs (665)
Education (346)
History (107)
Humor (99)
Instruments (11418)
Living History (56)
Lyrics (787)
Marching (1379)
Movies (335)
Museums (46)
Music Videos (129)
Musical Theatre (1831)
Musicology (264)
News and Media (197)
Organizations (127)
People (34)
Personal Pages (194)
Photography (136)
Radio (36)
Record Labels (2900)
Regional (11)
Resources (173)
Reviews (685)
Shopping (2935)
Songwriting (405)
Sound Files (1484)
Styles (31892)
Technology (96)
Television Shows (178)
Theory (174)
Trading (328)
Video Games (153)
Vocal (2952)
Weblogs (105)
Weddings (13)
Women in Music (1716)

Hierarchical clustering as a way to organize web pages

Finding association patterns in data …



Share your own customer images
Search inside this book

Pattern Recognition and Machine Learning (Information Science and Statistics) (Hardcover)
by Christopher M. Bishop (Author)
★★★★☆ (22 customer reviews)

List Price: ~~$74.95~~
Price: **$50.11** & this item ships for **FREE with Super Saver Shipping**. Details
You Save: $24.84 (33%)

Upgrade this book for $14.99 more, and you can read, search, and annotate every page online. See details
Availability: In Stock. Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Monday, October 29? Order it in the next 0 hours and 33 minutes, and choose **One-Day Shipping** at checkout. See details

58 used & new available from $43.59

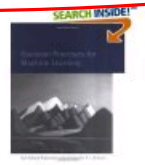
Better Together

Buy this book with The Elements of Statistical Learning by T. Hastie today!

Buy Together Today: $118.84
Buy both now!

Customers Who Bought This Item Also Bought

Pattern Classification (2nd Edition) by



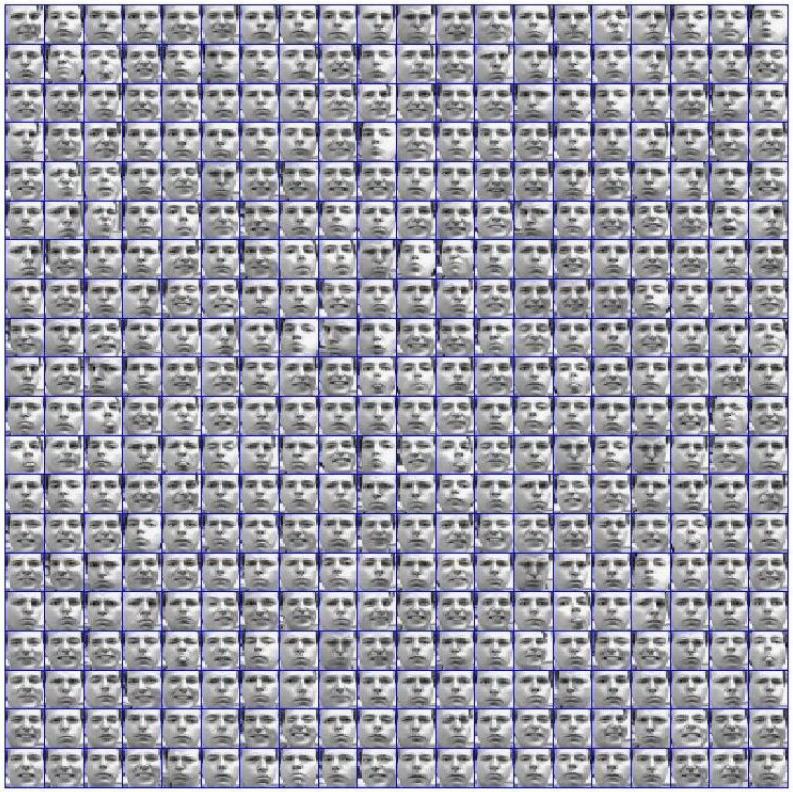
Gaussian Processes for Machine Learning by

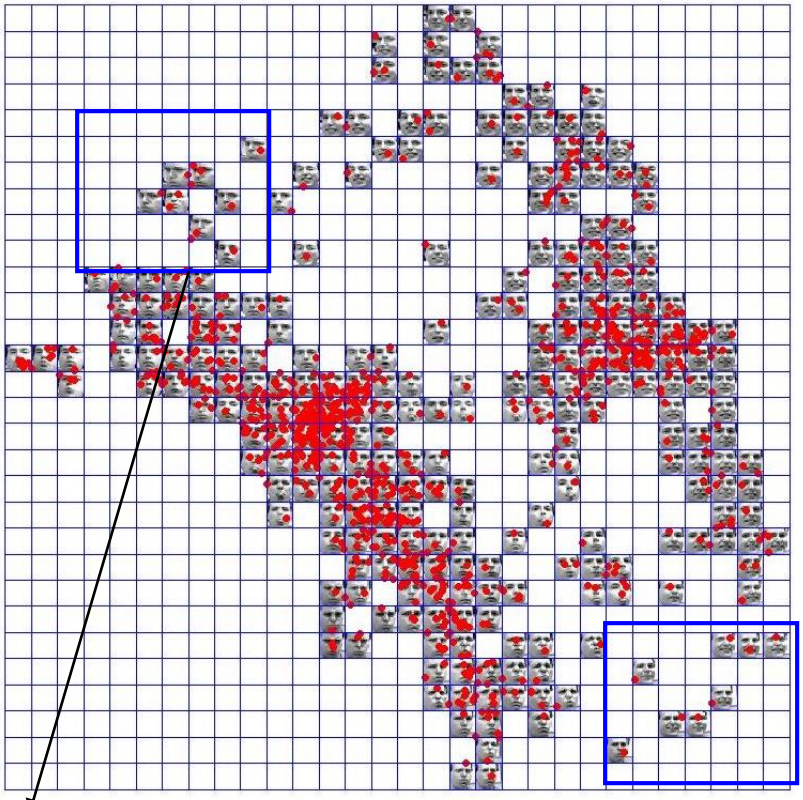
Data Mining by Ian H. Witten

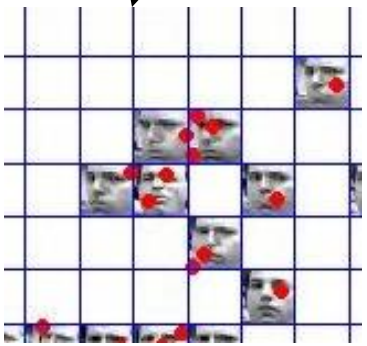
Machine Learning by Tom M. Mitchell

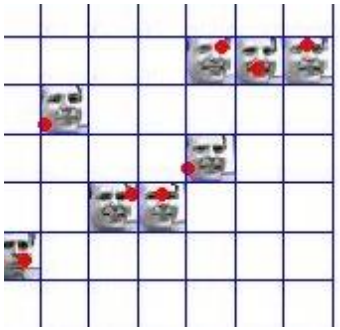
Computer Manual in MATLAB to Accompany

Dimension reduction



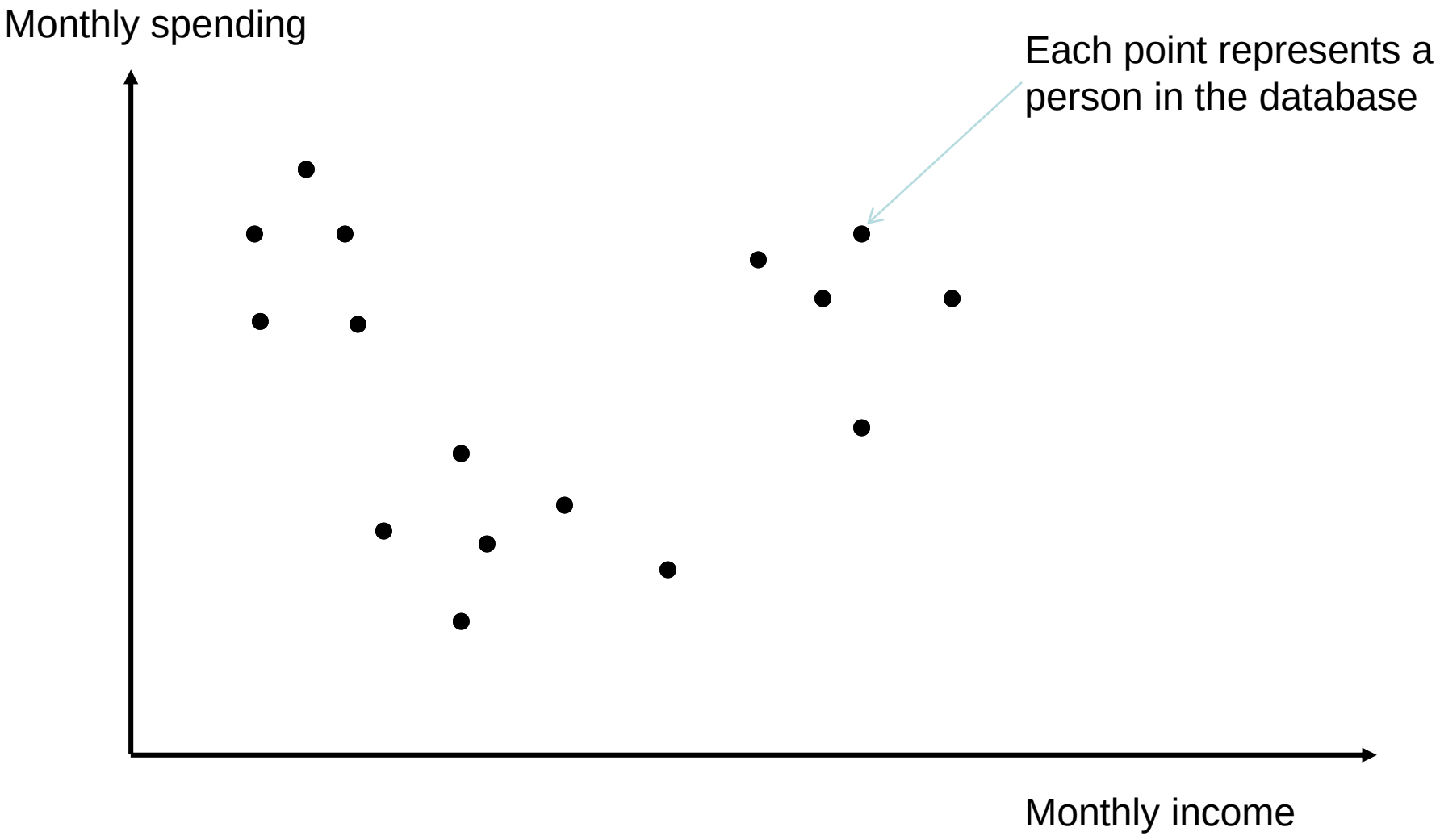




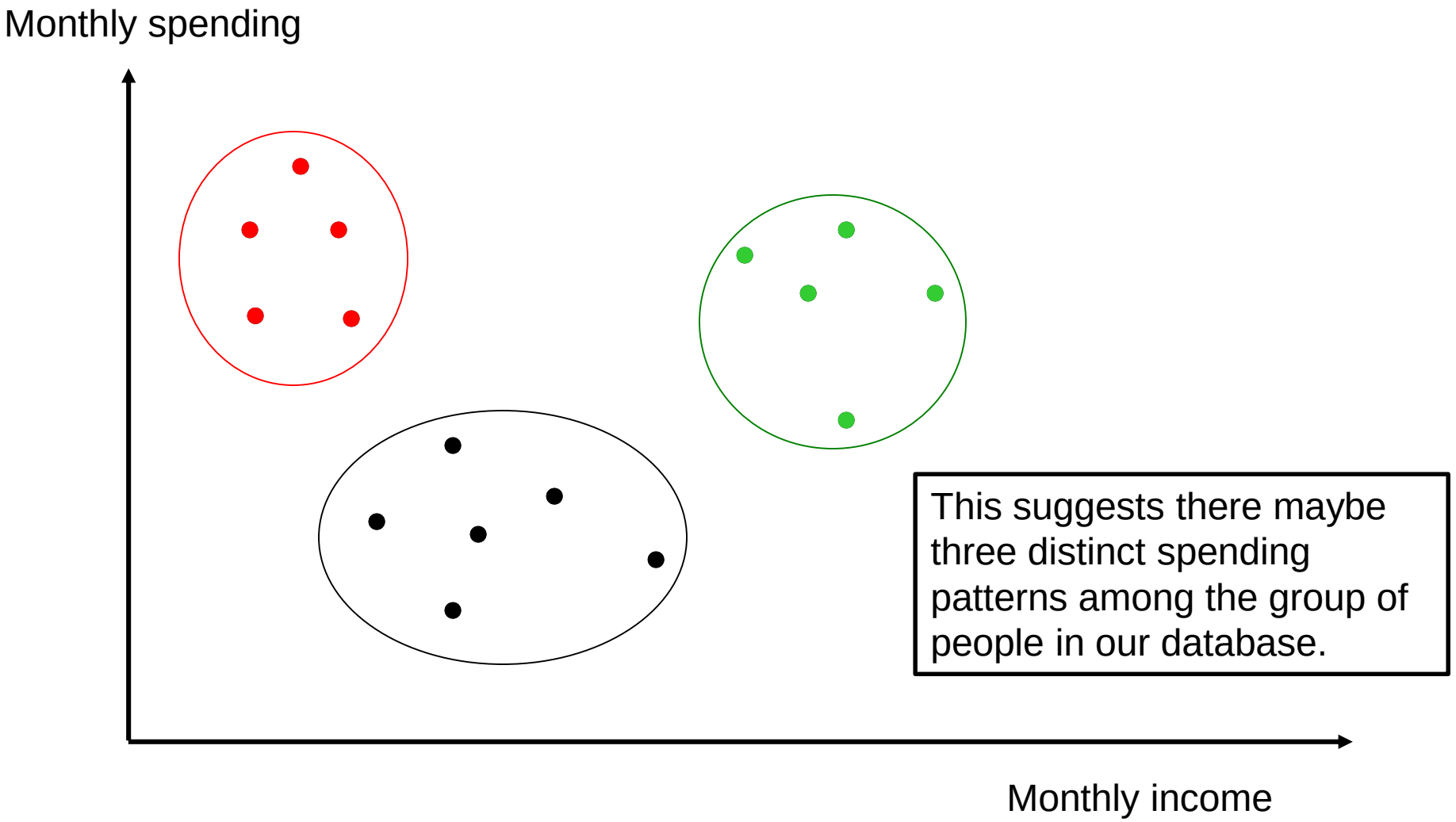


clustering

A hypothetical clustering example



A hypothetical clustering example



What is Clustering

- In general clustering can be viewed as an exploratory procedure for finding interesting subgroups in given data
- A large portion of the efforts focus on a special kind of clustering:
  - Group all given examples (i.e., ***exhaustive***) into disjoint clusters (i.e., ***partitional***), such that:
    - Examples within a cluster are (very) similar
    - Examples in different clusters are (very) different

Some example applications

- Information retrieval: cluster retrieved documents to present more organized and understandable results
- Consumer market analysis: cluster consumers into different interest groups so that marketing plans can be specifically designed for each individual group
- Image segmentation: decompose an image into regions with coherent color and texture
- Vector quantization for data (i.e., image) compression: group vectors into similar groups, and use group mean to represent group members
- Computational biology: group gene into co-expressed families based on their expression profile using different tissue samples and different experimental conditions

Image compression: Vector quantization

Group all pixels into self-similar groups, instead of storing all pixel values, store the means of each group

701,554 bytes → 127,292 bytes

Important components in clustering

- Distance/similarity measure
  - How to measure the similarity between two objects?
- Clustering algorithm
  - How to find the clusters based on the distance/similarity measures
- Evaluation of results
  - How do we know that our clustering result is good

Distance/similarity Measures

- One of the most important question in unsupervised learning, often more important than the choice of clustering algorithms
- What is similarity?
  - Similarity is hard to define, but we know it when we see it

  - More pragmatically, there are many mathematical definitions of distance/similarity

Common distance/similarity measures

- Euclidean distance $L_2(\mathbf{x}, \mathbf{x}')^2 = \sum_{i=1}^{d} (x_i - x_i')^2$

  Straight line distance >=0

- Manhattan distance

$$L_1(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d} |x_i - x_i'|$$

  City block distance >=0

- Cosine similarity

$$cos(\mathbf{x}, \mathbf{x}') = \frac{<\mathbf{x} \cdot \mathbf{x}'>}{|\mathbf{x}| \cdot |\mathbf{x}'|}$$

  Angle between two vectors, commonly used for measuring document similarity

- More flexible measures: $D(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{d} w_i (x_i - x_i')^2}$

  Scale each feature differently using w_i's

one can learn the appropriate weights given user guidance

Note: We can always transform between distance and similarity using a monotonically decreasing function, for example $e^{-\alpha D(x_1, x_2)^2}$

How to decide which to use?

- It is application dependent
- Consider your application domain, you need to ask questions such as:
  - What does it mean for two consumers to be similar to each other? Or for two genes to be similar to each other?
- For example, for text domain, we typically use cosine similarity
- This may or may not give you the answer you want depends on your existing knowledge of the domain
  - When domain knowledge is not enough, one could consider learning based approach

A learning approach

- Ideally we'd like to learn a distance function from user's inputs
  - Ask users to provide things like object A is similar to object B, dissimilar to object C
  - For example, if a user want to group the marbles based on their pattern

They don't belong together!

They belong together!

  - Learn a distance function to correctly reflect these relationships
    - E.g. weigh pattern-related features more than color features
  - This is a more advanced topic that we will not cover in this class, but nonetheless very important

- When we can not afford to learn a distance measure, and don’t have a clue about which distance function is appropriate, what should we do?

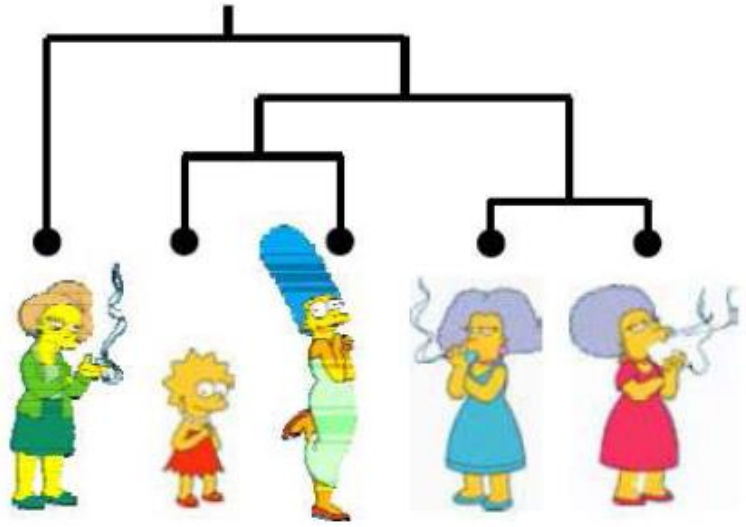
- Clustering is an exploratory procedure and it is important to explore – i.e., try different options and see what makes sense

Clustering

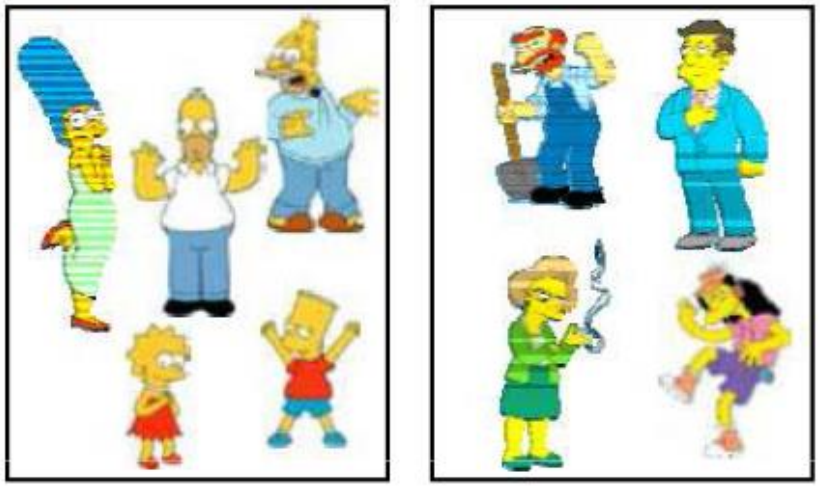
- Now we have briefly discussed distances or similarities among objects, how can we use them to group the objects into clusters?
- Many different approaches are available and they can be roughly categorized into two distinct types

Clustering algorithms

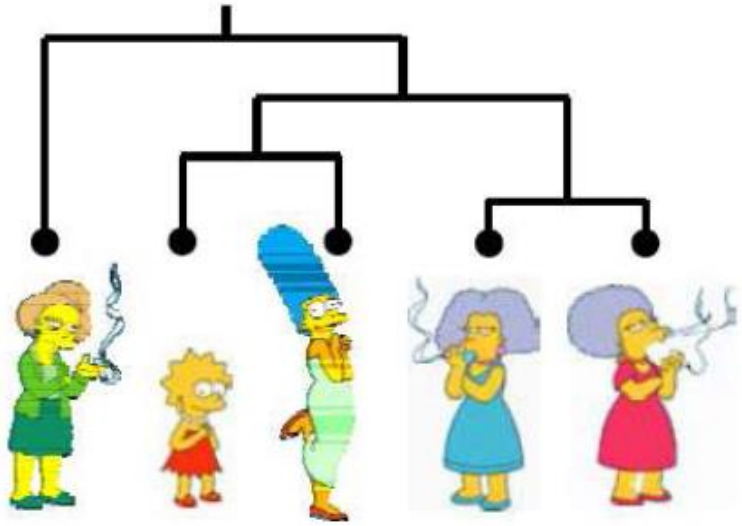
- Hierarchical algorithms
  - Bottom up – agglomerative
  - Top down – divisive



- Partition algorithms (Flat)
  - K-means
  - Mixture of Gaussian



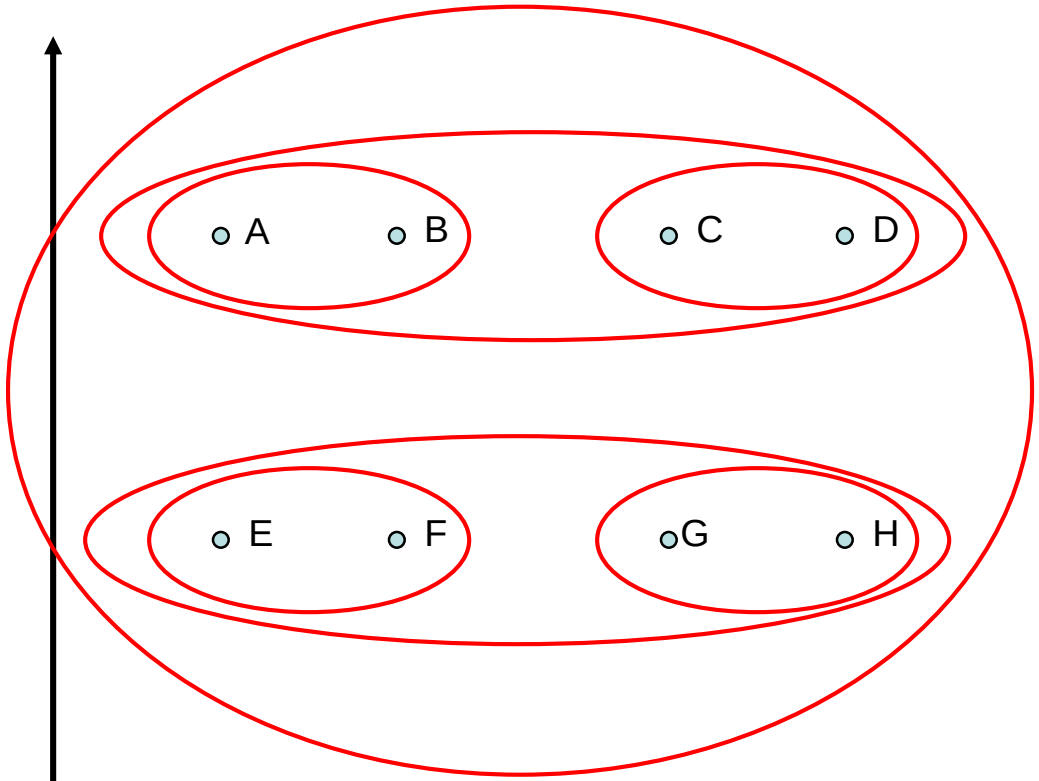
Hierarchical Clustering

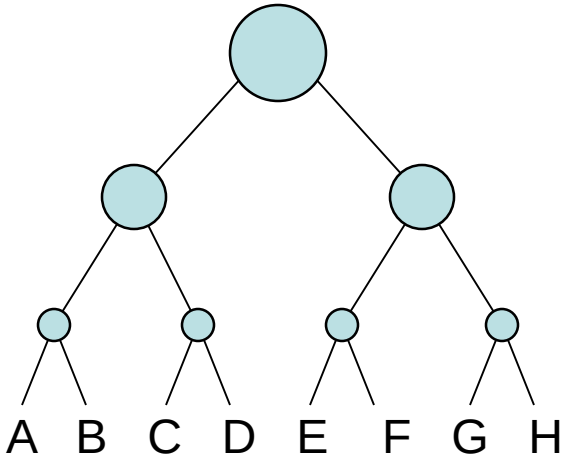


Hierarchical Agglomerative Clustering (HAC)

- Assumes a *distance function* for determining the similarity of two instances.
  - One can also assume a similarity function, and reverse some of the operations (e.g., minimum distance -> maximum similarity) in the algorithm to make it work for similarities
- Starts with each object in a separate cluster and then repeatedly joins the two closest clusters until only one cluster is left

HAC Example





HAC Algorithm

Start with all objects in their own cluster.
Repeat until there is only one cluster:
    Among the current clusters, determine the two clusters, c_i and c_j, that are closest
    Replace c_i and c_j with a single cluster $c_i \cup c_j$

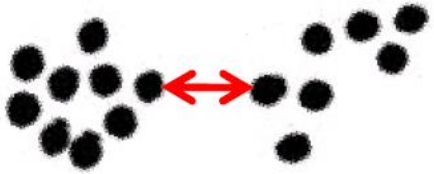
Problem: we assume a distance/similarity function that computes distance/similarity between examples, but here we also need to compute distance between clusters.

How?

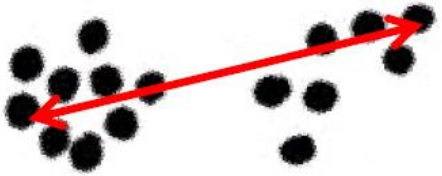
Closest Pair of Clusters

The distance between two clusters is defined as the distance between:

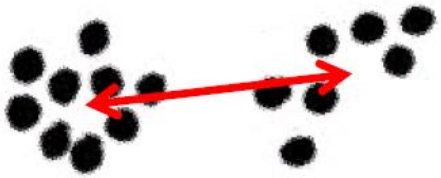
- Single-link
  - The nearest pair of points



- Complete-link
  - The furthest pair of points

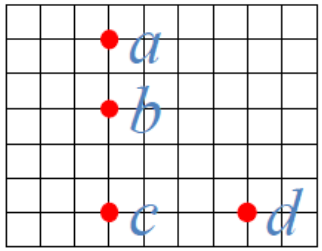


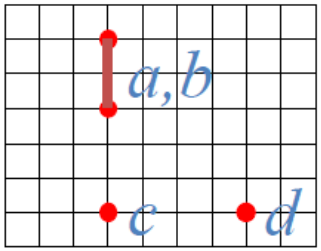
- Centroid
  - The center of gravity



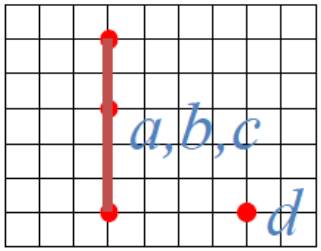
- Average-link
  - Average of all cross-cluster pairs
  - Most commonly used and most robust

Single Link Method

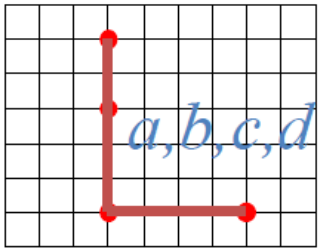




(1)



(2)



(3)

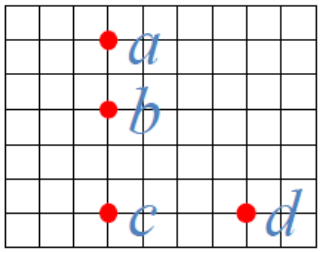
	b	c	d
a	2	5	6
b		3	5
c			4

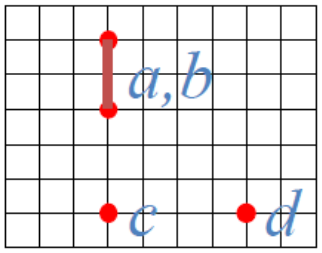
	b	c	d
a	2	5	6
b		3	5
c			4

	c	d
a, b	3	5
c		4

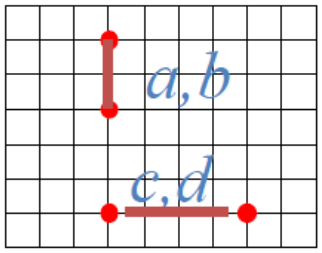
	d
a, b, c	4

Complete Link Method

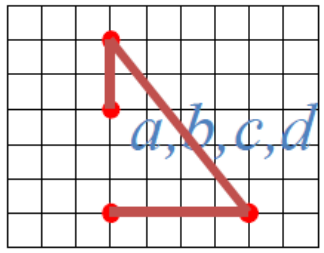




(1)



(2)



(3)

	b	c	d
a	2	5	6
b		3	5
c			4

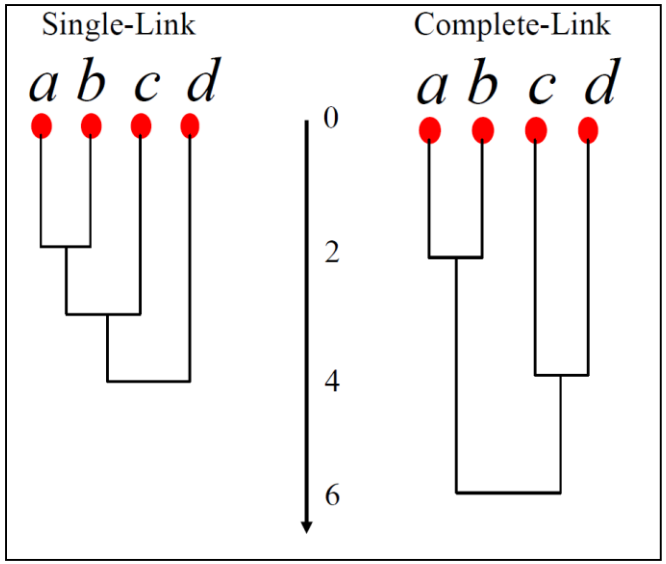
	b	c	d
a	2	5	6
b		3	5
c			4

	c	d
a,b	5	6
c		4

	c,d
a,b	6

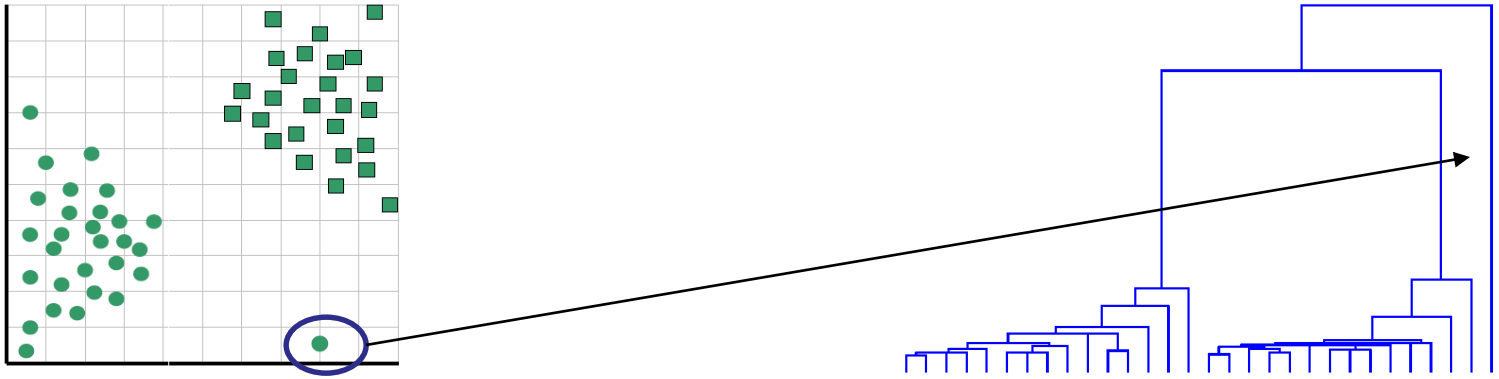
Visualization: Dendrogram

- Height of the joint = the distance between the two merge clusters
- The merge distance monotonically increases as we merge more and more for
  - Single, complete and average linkage methods
  - Not for the centroid method



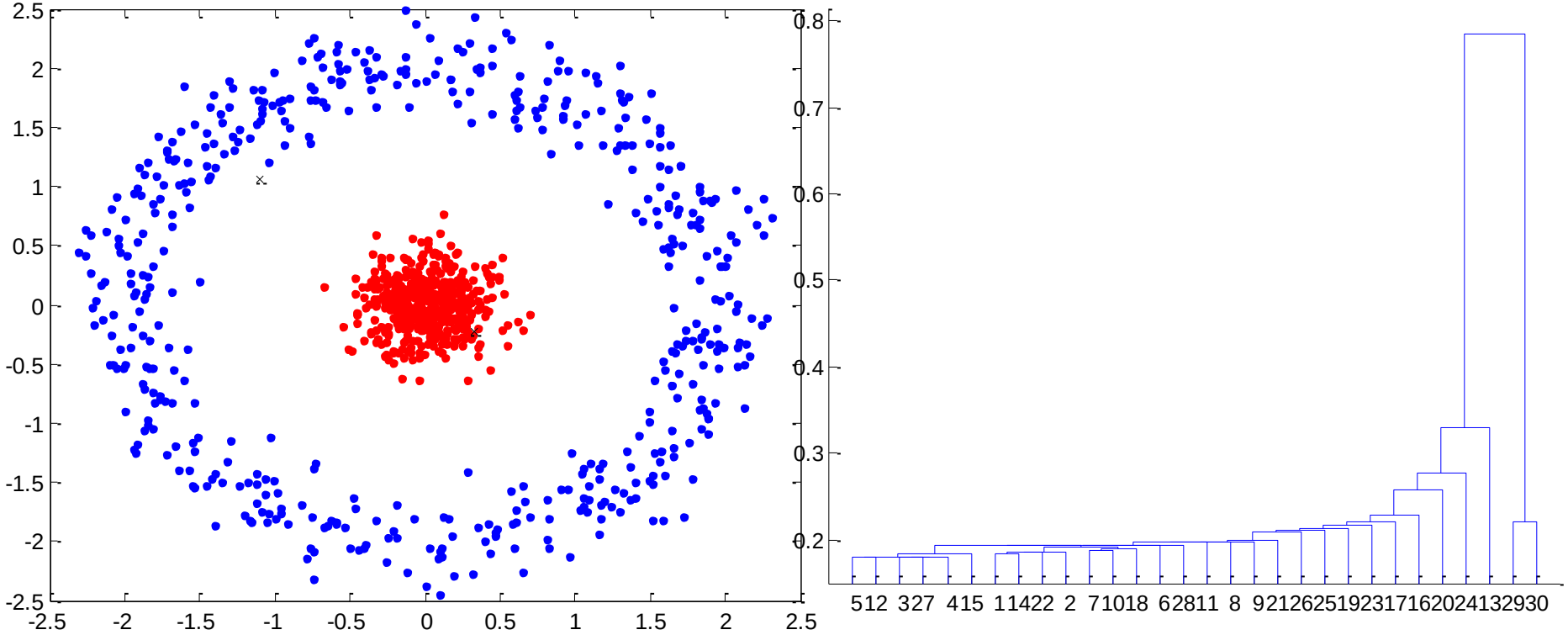
This example is shown upside down.

Interpreting Dendrogram

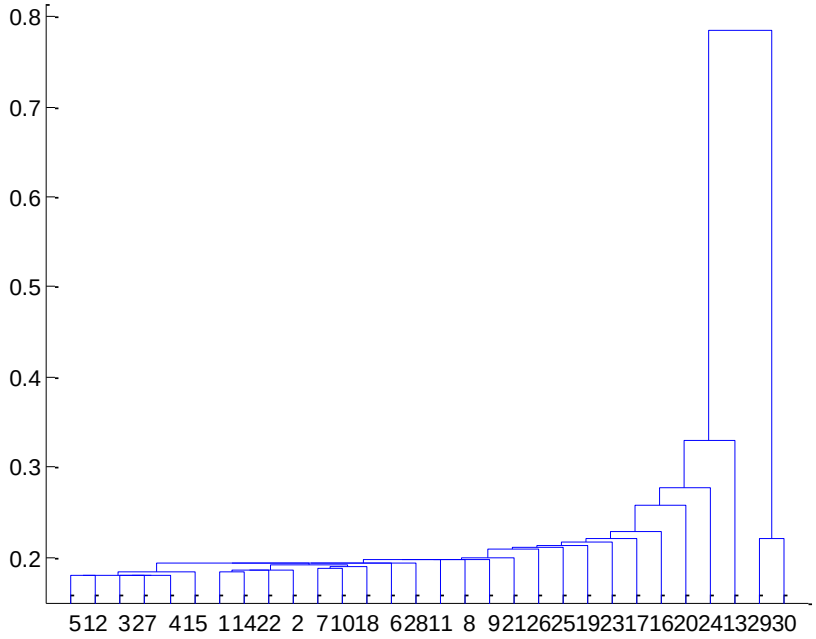


- Dendragram can be used to identify
  - The number of clusters in data
  - Well-formed clusters
  - Outliers

Single Link


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5
0.8
0.7
0.6
0.5
0.4
0.3
0.2
5 12 3 27 4 15 1 14 22 2 7 10 18 6 28 11 8 9 21 26 25 19 23 17 16 20 24 13 29 30

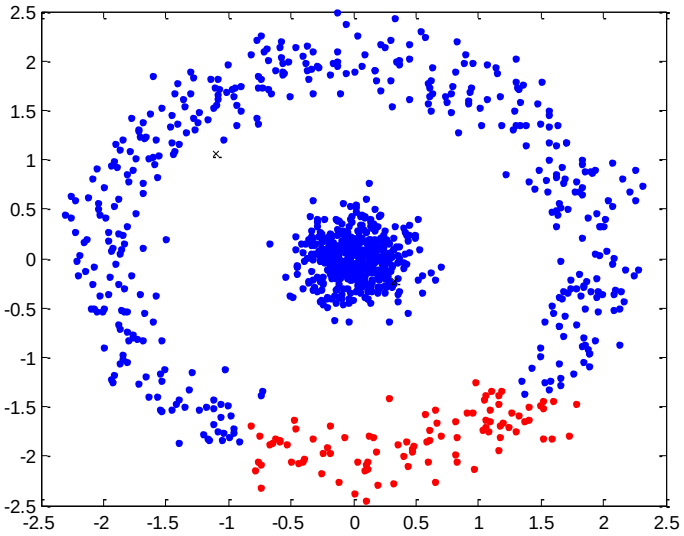
Single-link's chaining effect



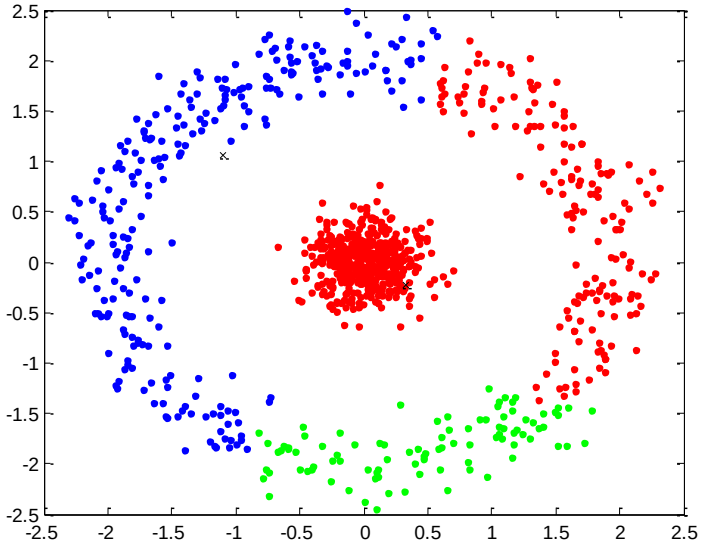
- Single link is famous for its chaining effect
- It can gradually adds more and more examples to the "chain"
- Create long straggling clusters

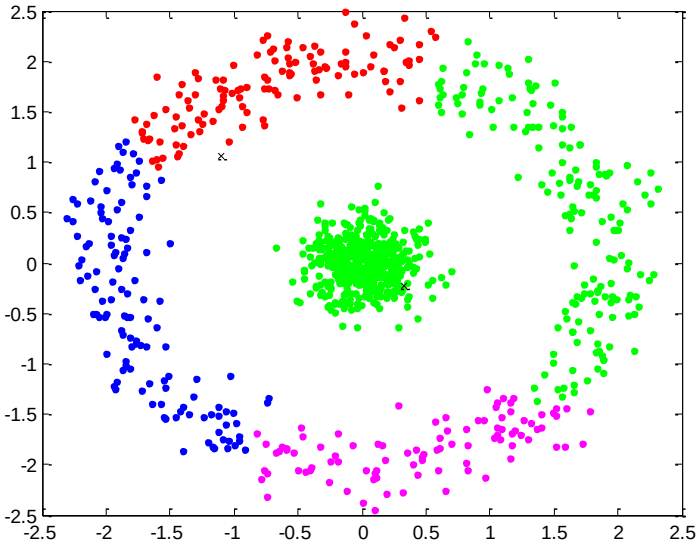
Complete Link

K=2

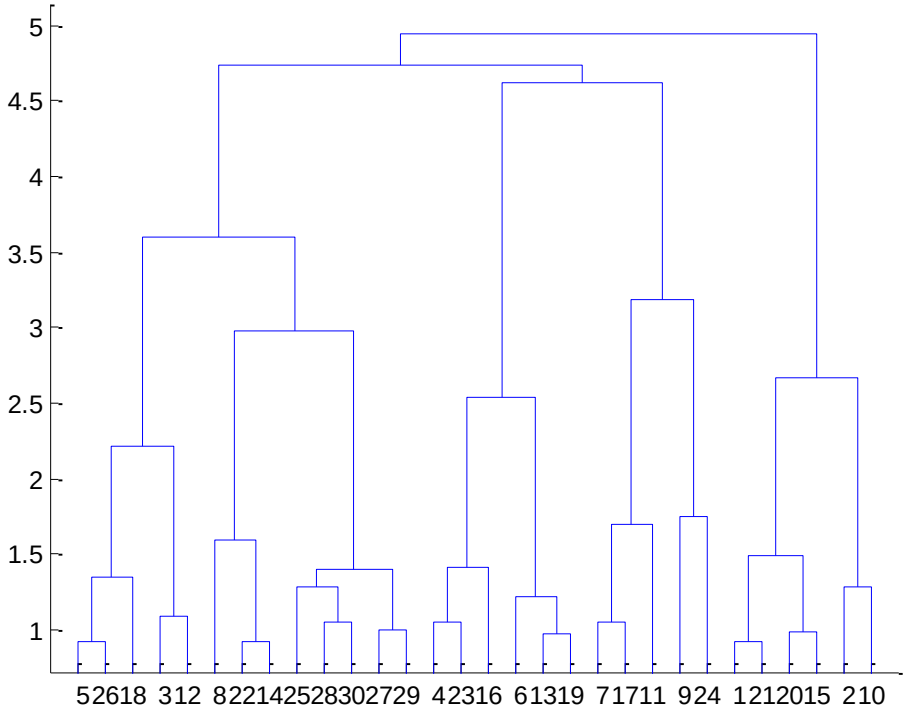


K=3

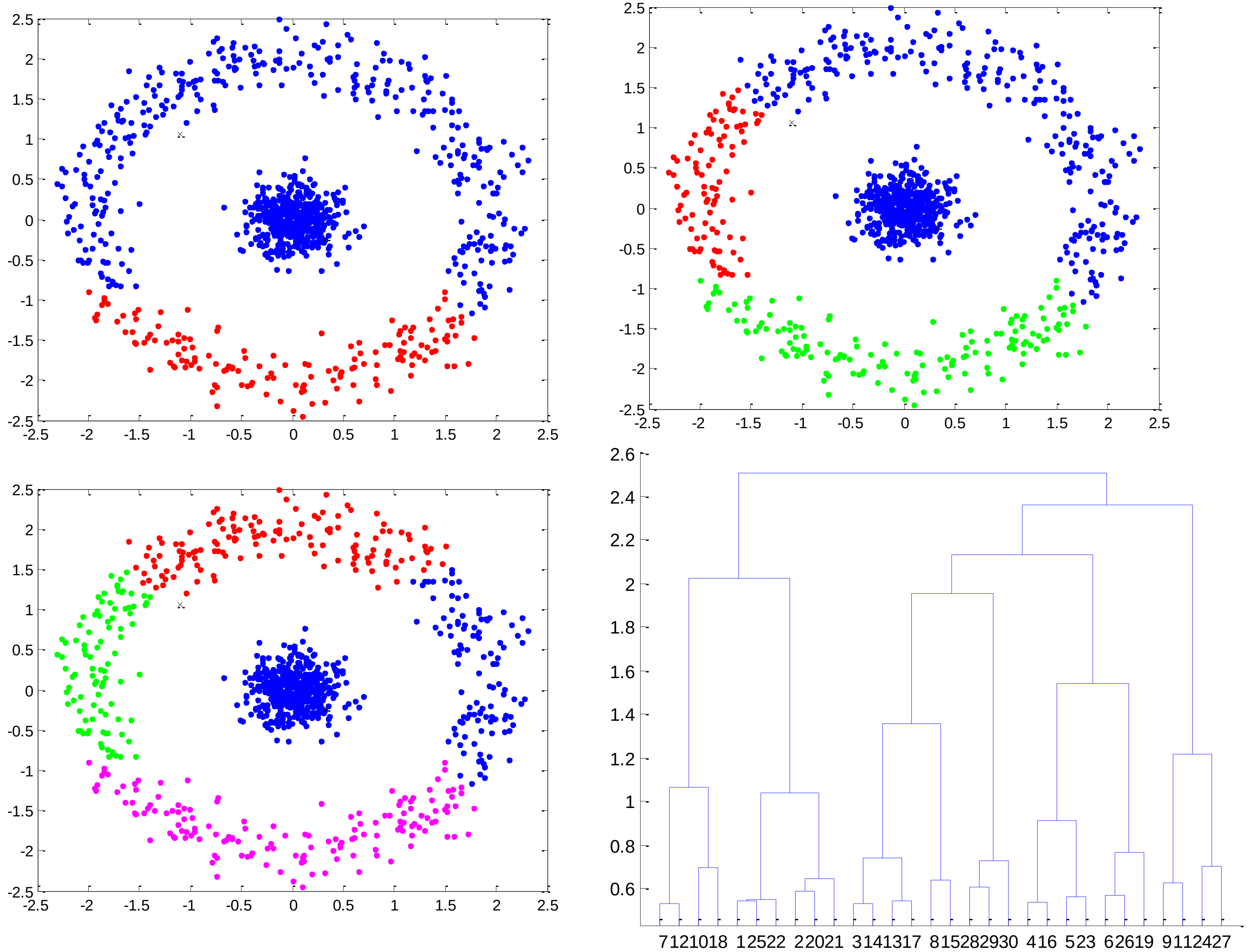




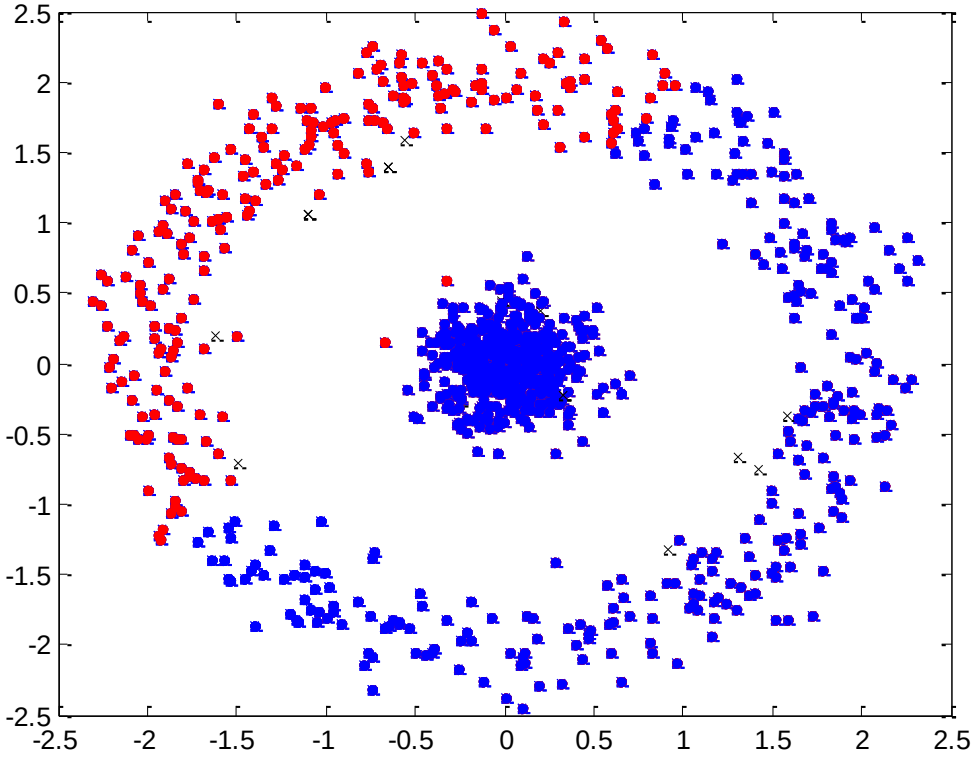
K=4



Average Link



K-means



Comments on HAC

- HAC is a convenient tool that often provides interesting views of a dataset
- Primarily HAC can be viewed as an intuitively appealing clustering procedure for data analysis/exploration
- We can create clusterings of different granularity by stopping at different levels of the dendrogram
- HAC often used together with visualization of the dendrogram to decide how many clusters exist in the data
- Different linkage methods (single, complete and average) often lead to different solutions

K-Means Clustering

Goal of Clustering

- Given a data set $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$
- We need to partition D into k disjoint clusters, such that
  - Examples are self-similar in the same cluster
- How do we quantify this?

Objective: Sum of Squared Errors

- Given a partition of the data into k clusters, we can compute the center (i.e., mean, center of mass) of each cluster

$$\mu_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$$

- For a well formed cluster, its points should be close to its center. We measure this with sum of squared error (SSE), and formulate our objective to find a partition $\mathbb{C}^*$ that minimizes sum of squared error:

$$\mathbb{C}^* = \underset{\mathbb{C}=\{C_1,\ldots,C_k\}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} \|\boldsymbol{x} - \boldsymbol{\mu}_i\|^2$$

- This is a combinatorial optimization problem
  - NP-hard
  - K-Means finds a local optimal solution using an iterative approach

Basic idea

- We assume the number of desired clusters, *k*, is given
- Randomly choose *k* examples as *seeds*, one for each cluster.
- Form initial clusters based on these seeds.
- Iterate by repeatedly re-allocating instances to different clusters to improve the overall clustering.
- Stop when clustering converges or after a fixed number of iterations.

The Standard K-means algorithm

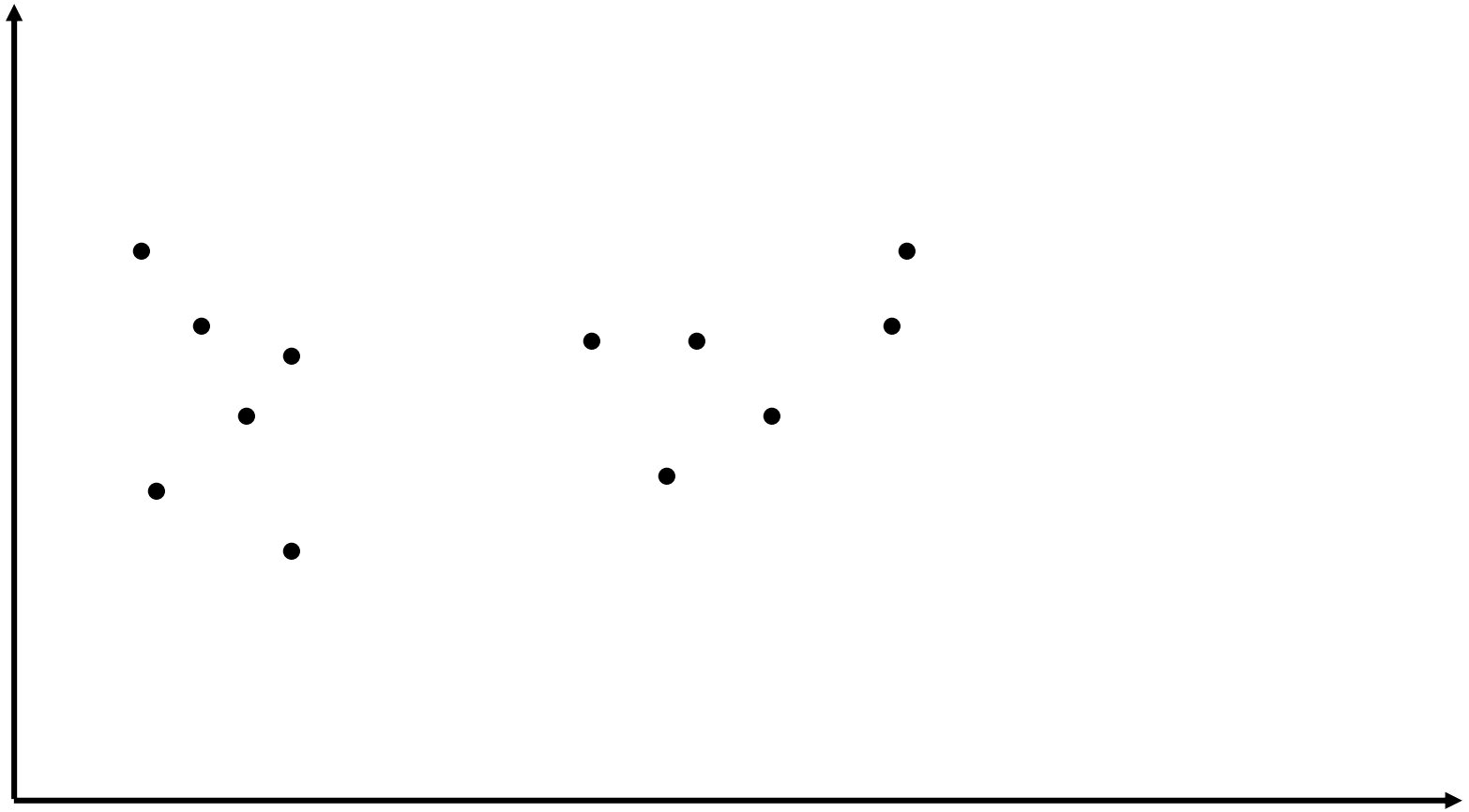
Input: $D = \{x_1\ x_2\ \ldots x_n\}$ and desired number of clusters k

Output: a partition of D into k disjoint clusters $c_1\ \ldots\ c_k$ (s.t. D = $c_1 \cup c_2 \cup \cdots \cup c_k$)

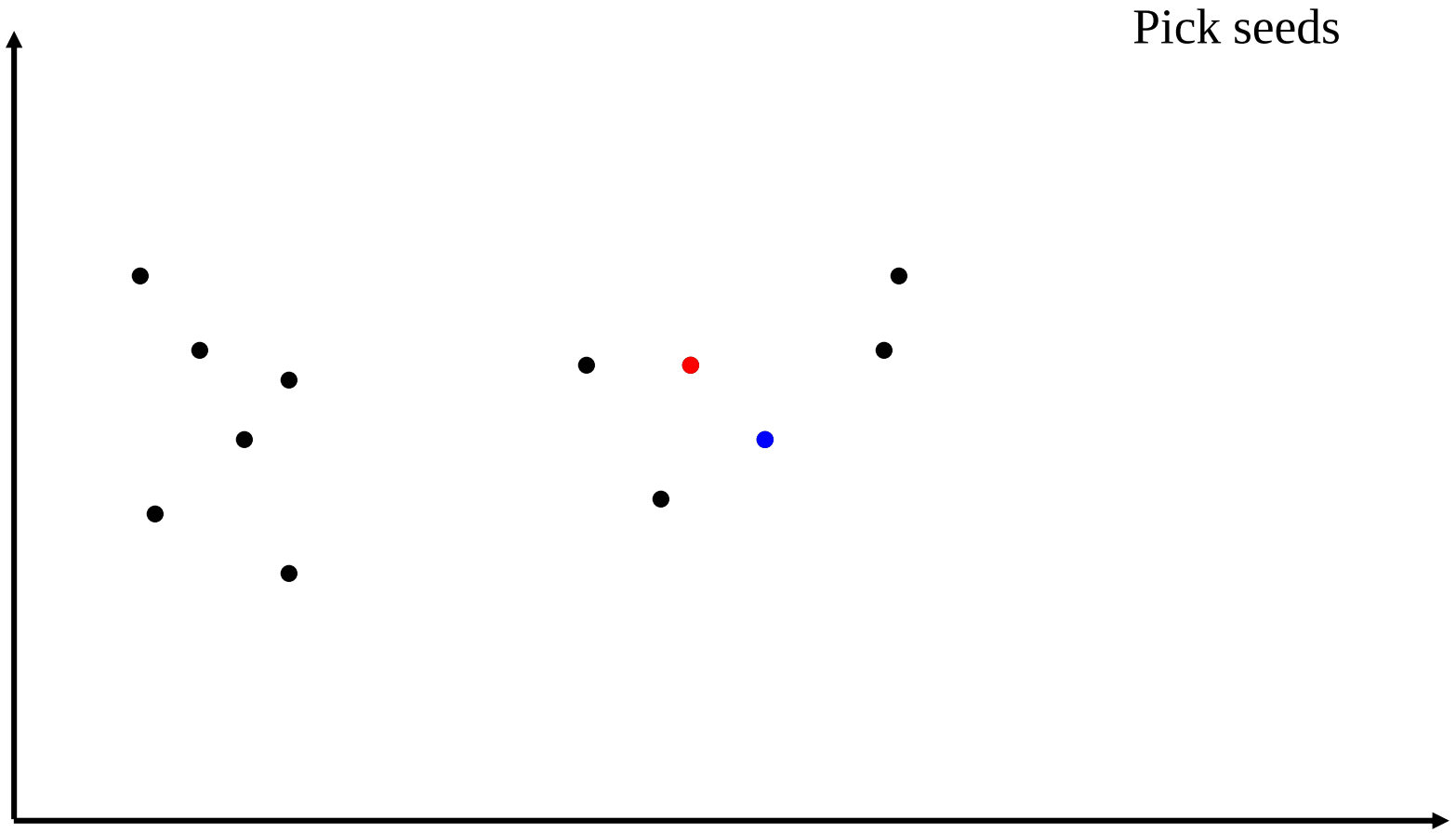
Let $\boldsymbol{d}$ be the distance function between examples

1. Select k random samples from D as centers $\{\mu_1\ \ldots \mu_k\}$ //**Initialization**
2. Do
3. for each example $\boldsymbol{x}_i$,
4. assign $\boldsymbol{x}_i$ to $\boldsymbol{c}_j$ such that $\boldsymbol{d}(\boldsymbol{\mu}_j, \boldsymbol{x}_1)$ is minimized // **the Assignment step**
5. for each cluster j, update cluster center
6. $$\mu_j = \frac{1}{|c_j|}\sum_{\mathbf{x}\in c_j}\mathbf{x}$$ // **the update step**
7. Until convergence

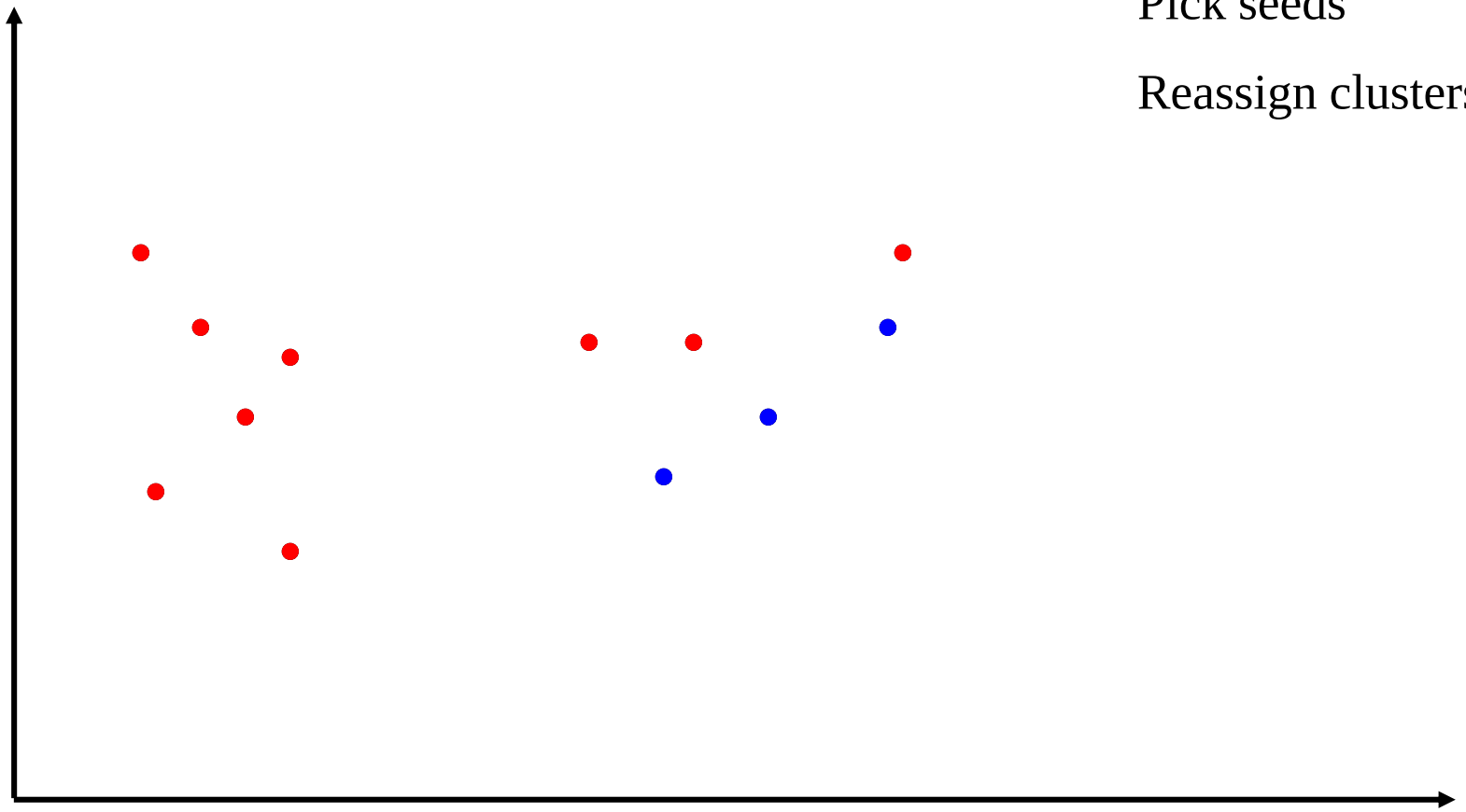
K-Means Example (K=2)



K-Means Example (K=2)



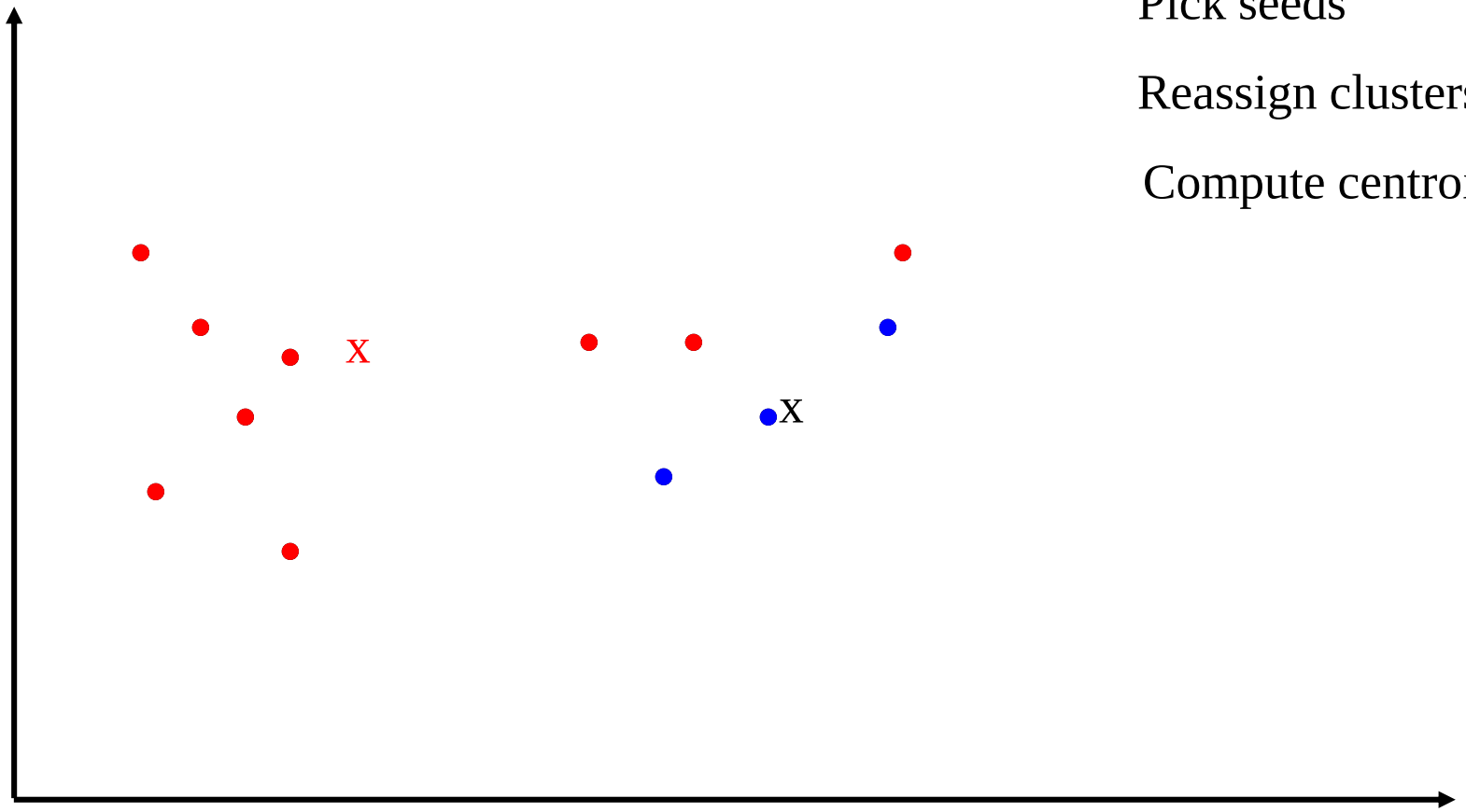
K-Means Example (K=2)



Pick seeds

Reassign clusters

K-Means Example (K=2)

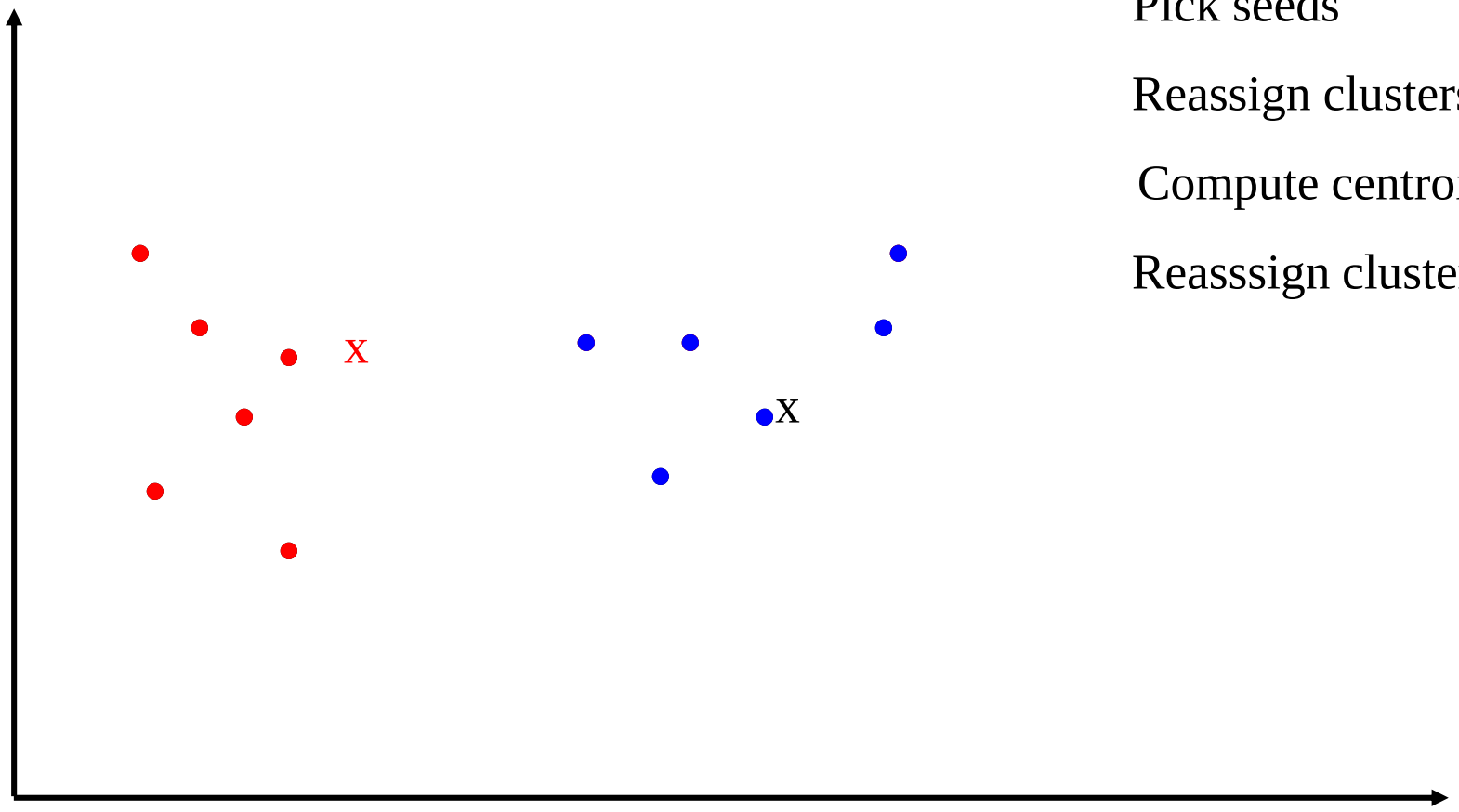


Pick seeds

Reassign clusters

Compute centroids

K-Means Example (K=2)



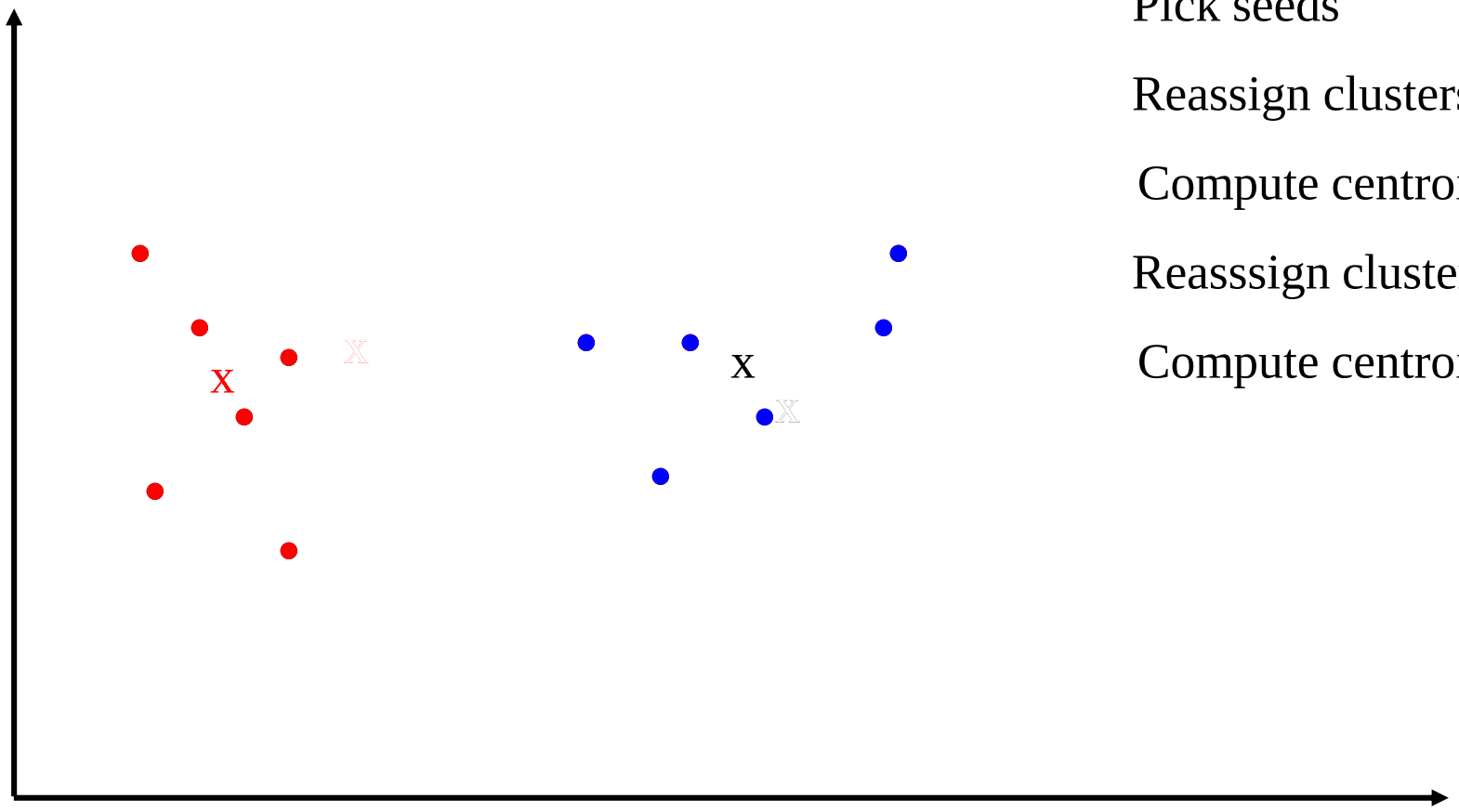
Pick seeds

Reassign clusters

Compute centroids

Reasssign clusters

K-Means Example (K=2)



Pick seeds

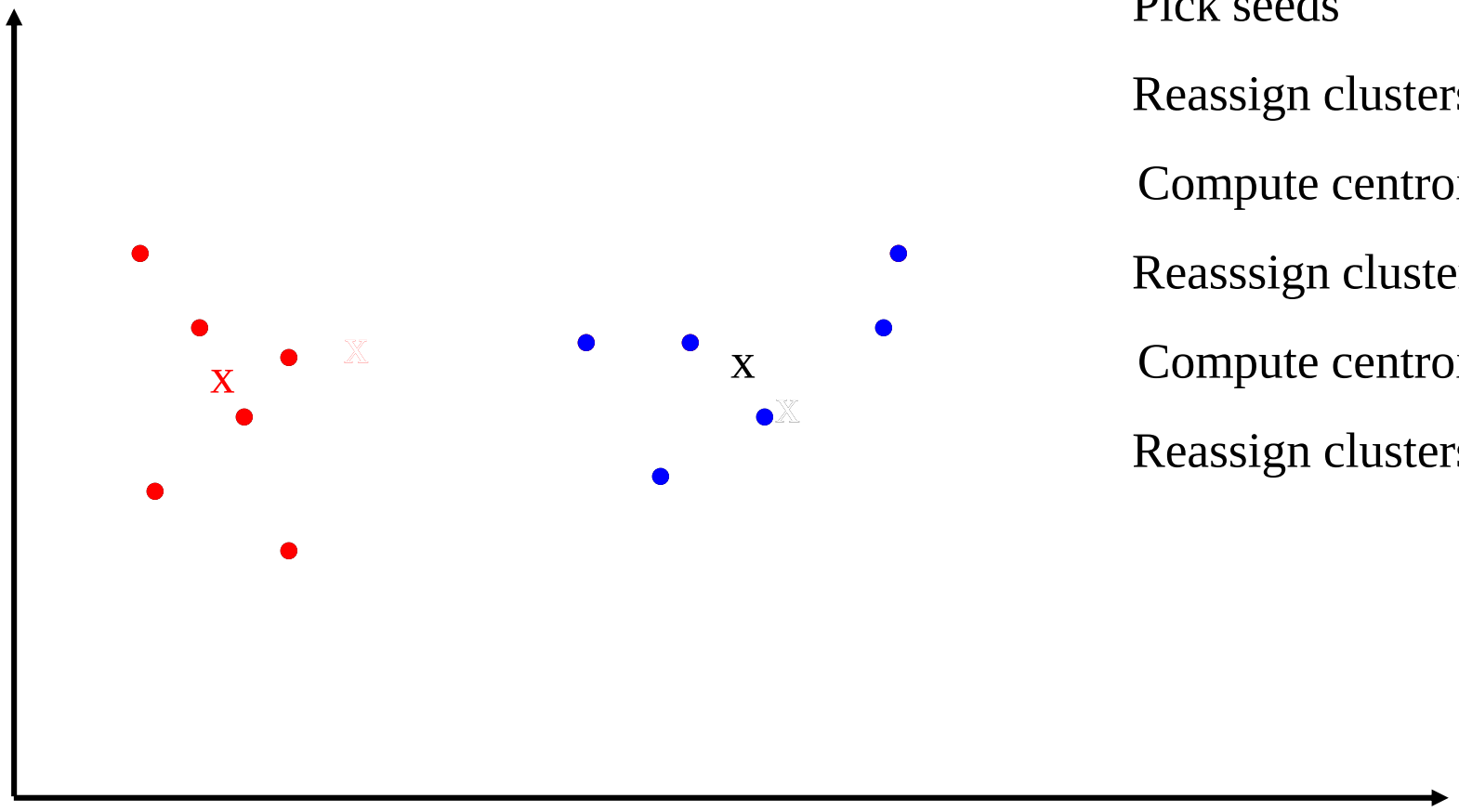
Reassign clusters

Compute centroids

Reasssign clusters

Compute centroids

K-Means Example (K=2)



Pick seeds

Reassign clusters

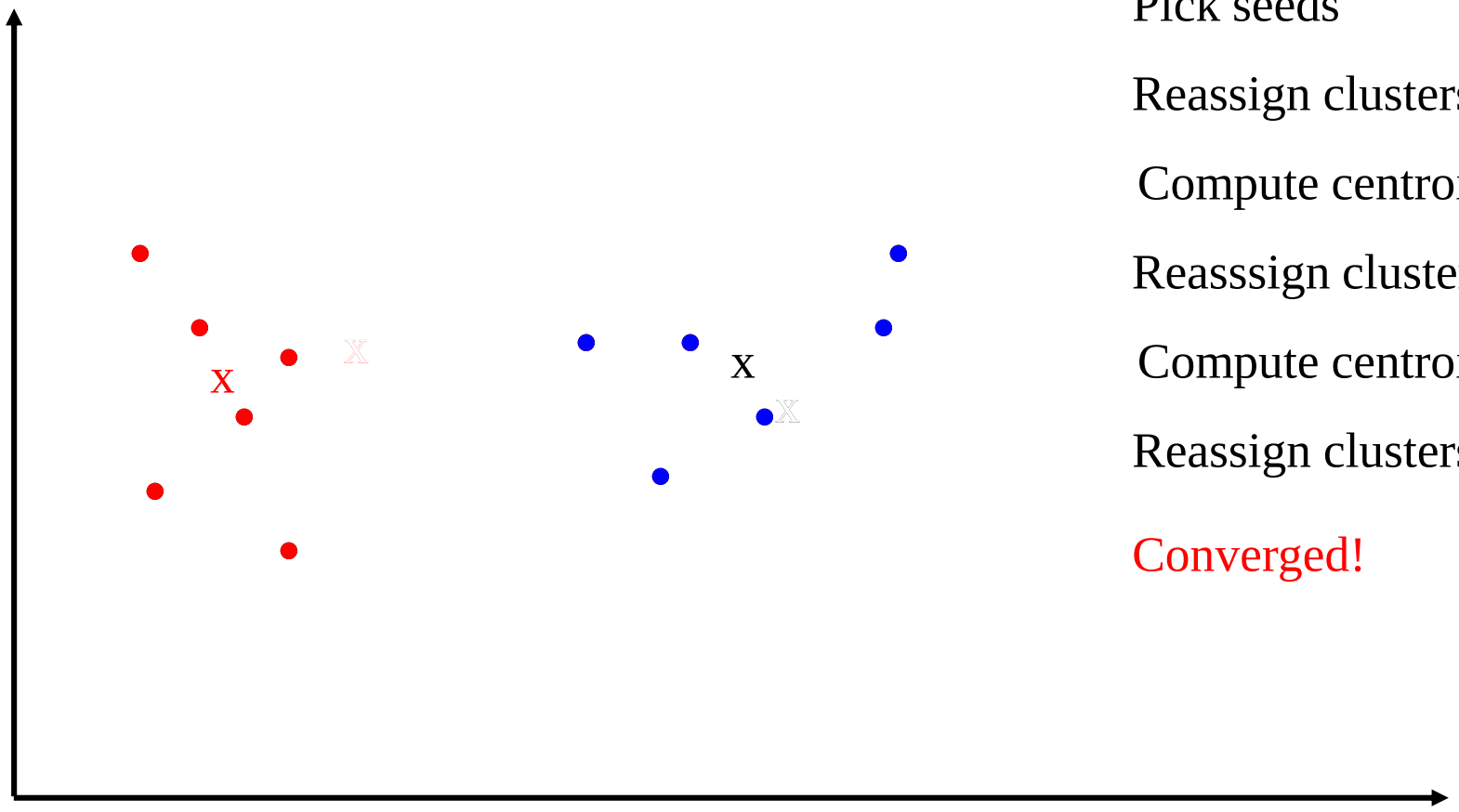
Compute centroids

Reasssign clusters

Compute centroids

Reassign clusters

K-Means Example (K=2)



Pick seeds

Reassign clusters

Compute centroids

Reasssign clusters

Compute centroids

Reassign clusters

Converged!

Monotonicity of K-means

- **Monotonicity Property:** Each iteration of K-means strictly decreases the SSE until convergence
- The following lemma is key to the proof:
  - **Lemma:** Given a finite set C of data points the value of μ that minimizes the SSE:

$$J = \sum_{\mathrm{X} \in C} \|\mathrm{x} - \mu\|^2$$

is:

$$\mu = \frac{1}{|C|} \sum_{\mathbf{x} \in C} \mathbf{x}$$

Proof of monoticity

- Given a current set of clusters with their means, the SSE is given by :

$$J_e = \sum_{i=1}^{k} \sum_{\mathrm{x} \in C_i} \|\mathrm{x} - \mu_i\|^2$$

- Consider the assignment step:
  - Since each point is only reassigned if it is closer to some other cluster than its current cluster, so we know the reassignment step will only decrease SSE
- Consider the re-center step:
  - From our lemma we know the updated μ_i minimizes the SSE of c_i, which implies that the resulting SSE again is decreased.
- Combine the above two, we know K-means always decreases the SSE

K-means properties

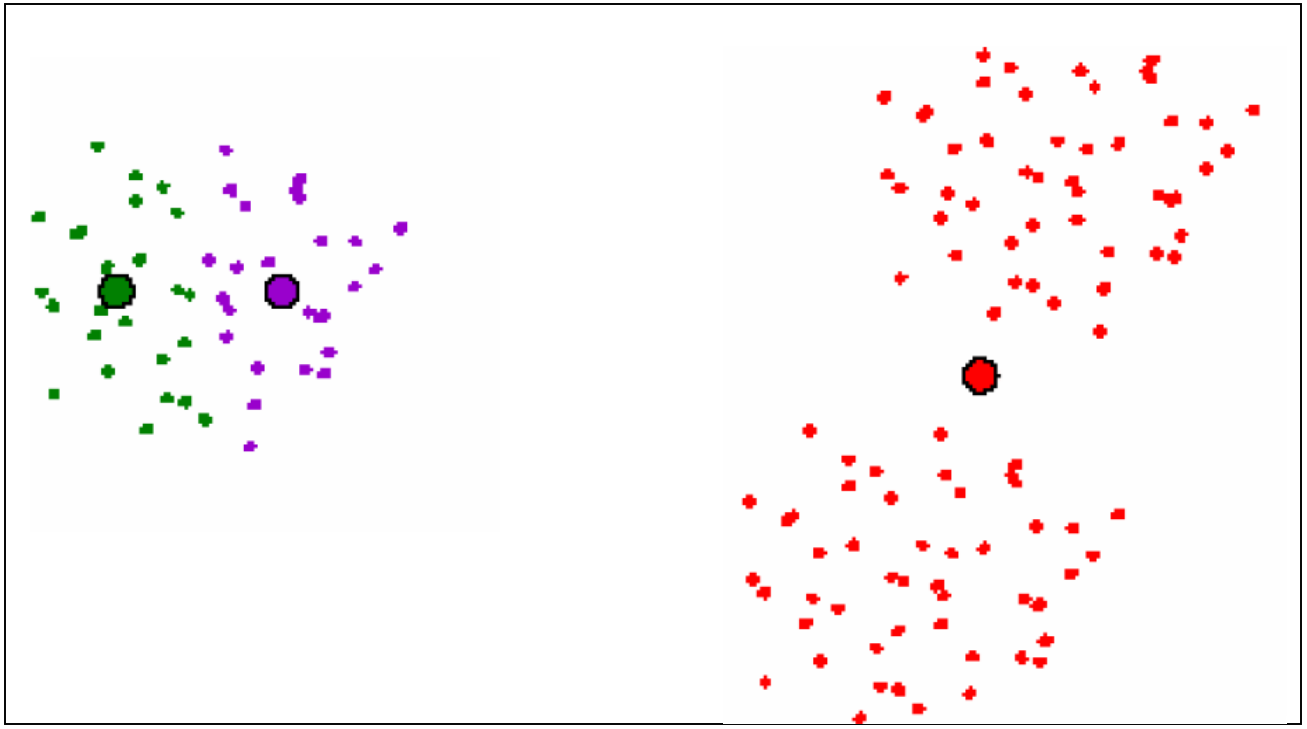
Good news!

- K-means always converges in a finite number of steps
  - Typically converges very fast (in fewer iterations than $\mathrm{n} = |D|$)
- Time complexity:
  - Assume computing distance between two instances is $O(d)$ where d is the dimensionality of the vectors.
  - Reassigning clusters: $O(kn)$ distance computations, or $O(knd)$.
  - Computing centers: Each instance vector gets added once to some center: $O(nd)$.
  - Assume these two steps are each done once for l iterations: $O(lknd)$.
  - Linear in all relevant factors, assuming a fixed number of iterations

More Comments

Bad news!

- Highly sensitive to the initial seeds



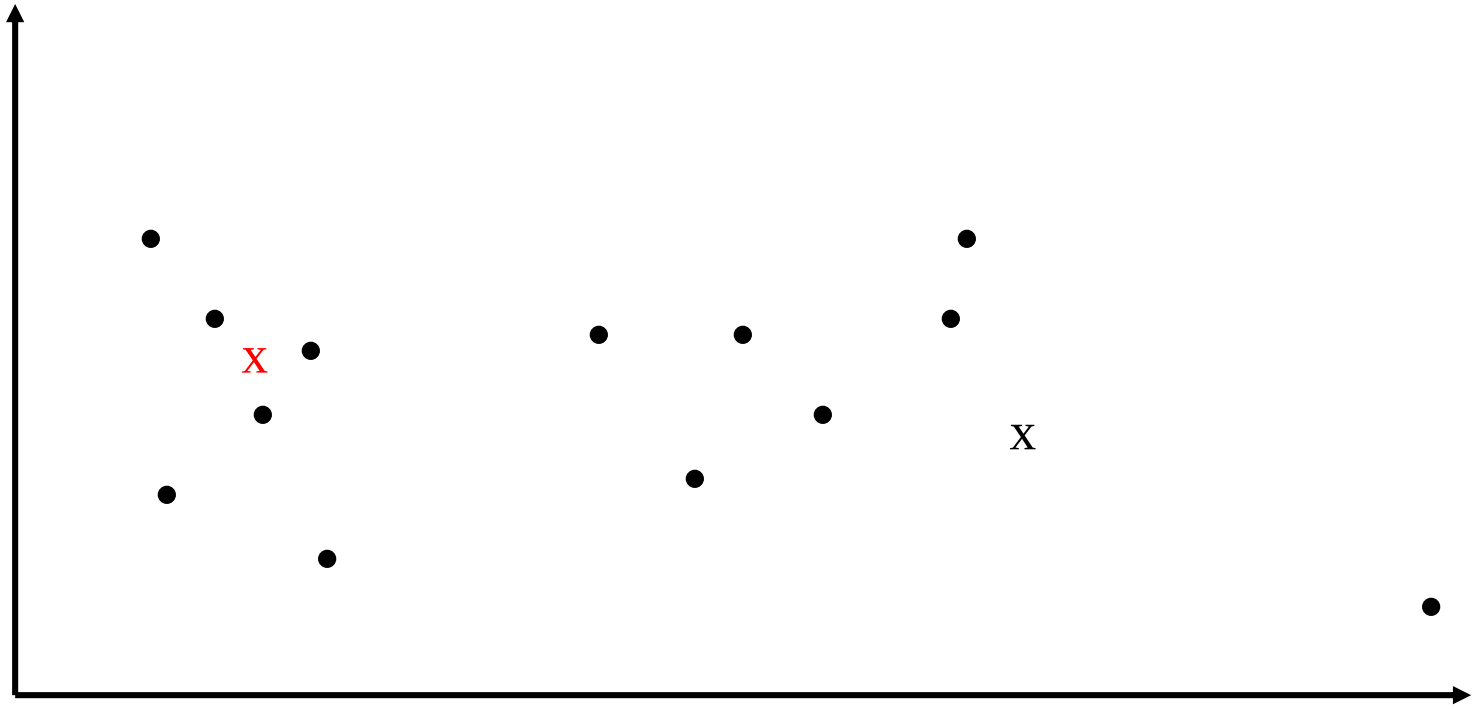
- This is because SSE has many local minimal solutions, i.e., solutions that can not be improved by local reassignments of any particular points

Solutions

- Run multiple trials and choose the one with the best SSE
  - This is typically done in practice
- Heuristics: try to choose initial centers to be far apart
  - Using furthest-first traversal
    - Start with a random initial center
    - Set the second center to be furthest from the first center
    - The third center to be furthest from the first two centers
    - …

Even more comments

- K-Means is exhaustive:
  - Cluster every data point, no notion of outlier
- Outliers may cause problems, why?



K-medoids

- Outliers strongly impact the cluster centers
- K-medoids – use the medoid for each cluster, i.e., the data point that is closest to other points in the cluster

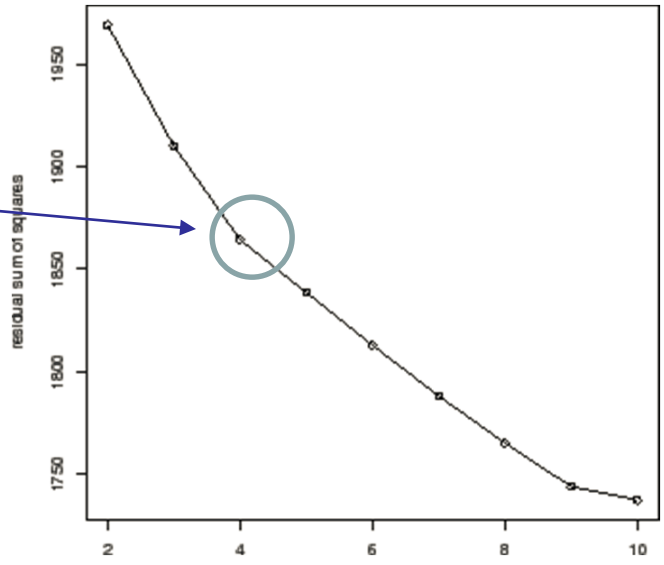
$$\mu = \frac{1}{|C|} \sum_{\mathbf{x} \in C} \mathbf{x}$$

$$\mu = \underset{\mathrm{x} \in C}{\operatorname{argmin}} \sum_{\mathrm{z} \in C} \|\mathrm{x} - \mathrm{z}\|^2$$

- K-medoids is computationally more expensive but more robust to outliers

Deciding ***k*** – a model selection problem

- What if we don't know how many clusters there are in the data?
- Can we use SSE to decide k by choosing k that gives the smallest SSE?
  - We will always favor larger k values
- Any quick solutions?
  - heuristic: find the knee



Mixture of Gaussians: Expectation Maximization

Hard vs. Soft Clustering

- Kmeans performs what we call hard clustering
  - Data point is deterministically assigned to one and only one cluster
- Soft-clustering:
  - Data points are assigned to clusters with certain probabilities
  - Typically assume some probabilistic model to represent the clusters

Soft clustering: Gaussian Mixture Modeling

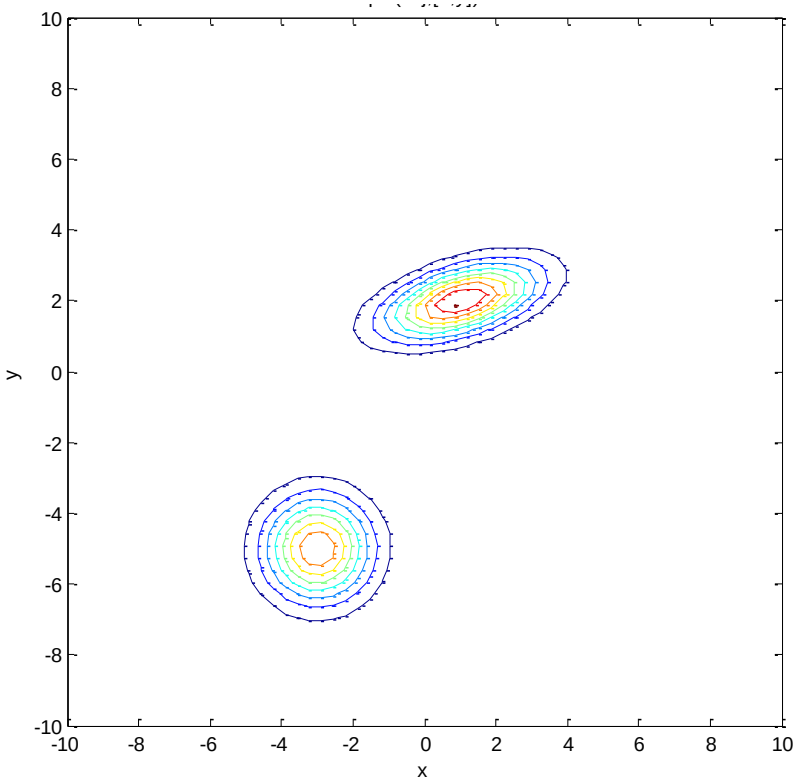
- Given data $\{x^1, \ldots, x^m\}$, assume k clusters
- Assume each cluster is generated from a Gaussian
- The proportion of the clusters are decided by $p(y = 1), \ldots, p(y = k)$

Example: two Gaussians

$$p(y = 1) = p(y = 2) = 0.5$$

$$\mu_1 = [1,2], \Sigma_1 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 5 \end{bmatrix}$$

$$\mu_2 = [-3, -5], \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Soft clustering: Gaussian Mixture Modeling

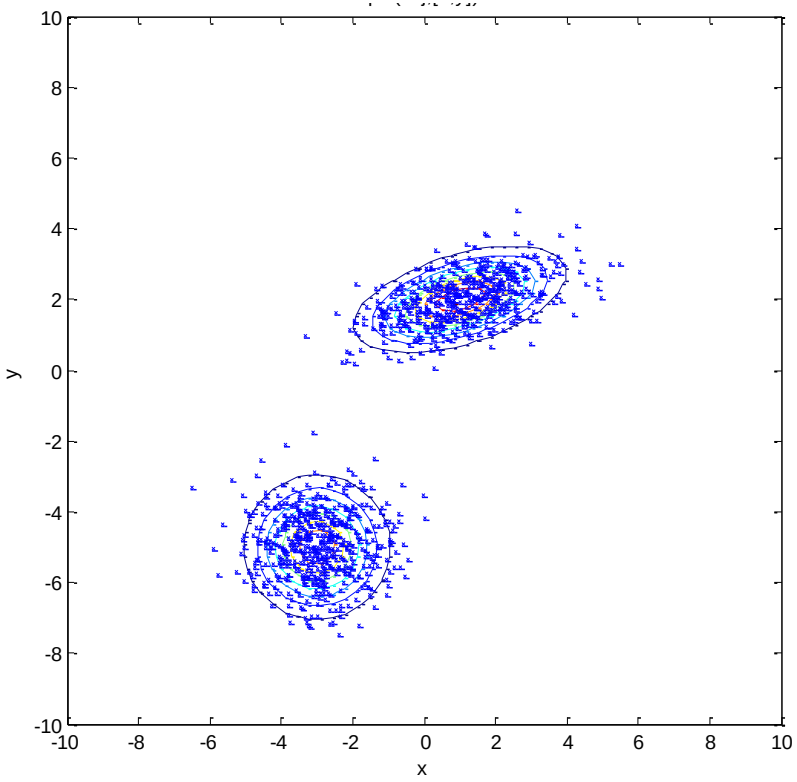
- Given data $\{x^1, \ldots, x^m\}$, assume k clusters
- Assume each cluster is generated from a Gaussian
- The proportion of the clusters are decided by $p(y = 1), \ldots, p(y = k)$

Example: sample points from the two Gaussians

$$p(y = 1) = p(y = 2) = 0.5$$

$$\mu_1 = [1,2], \Sigma_1 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 5 \end{bmatrix}$$

$$\mu_2 = [-3, -5], \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Soft clustering: Gaussian Mixture Modeling

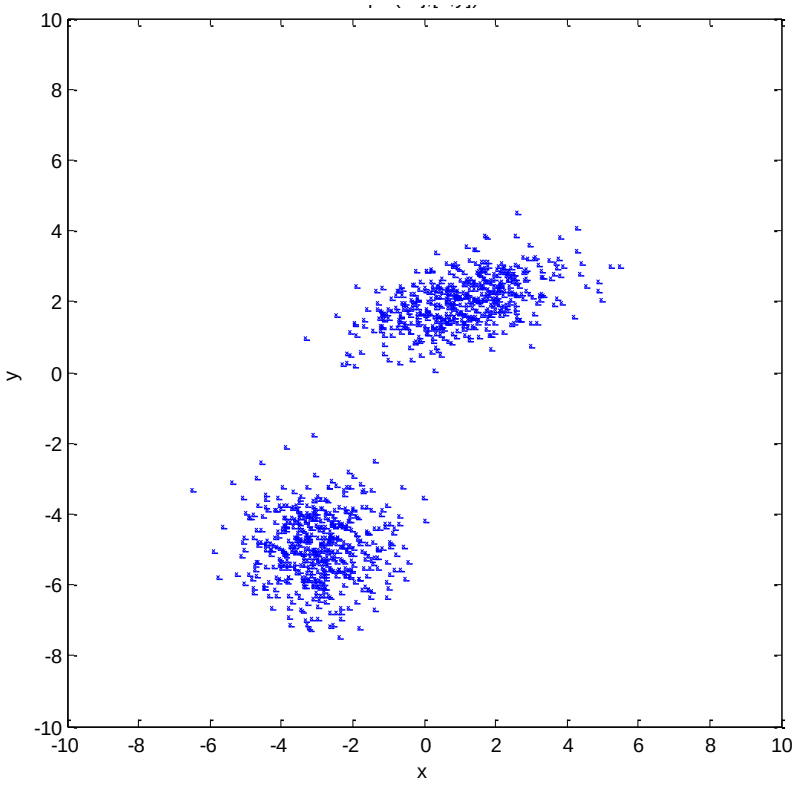
- Given data $\{x^1, \ldots, x^m\}$, assume k clusters
- Assume each cluster is generated from a Gaussian
- The proportion of the clusters are decided by $p(y = 1), \ldots, p(y = k)$

Example: observe the points, need to estimate the Gaussians

$$p(y = 1) = p(y = 2) = 0.5$$

$$\mu_1 = [1,2], \Sigma_1 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 5 \end{bmatrix}$$

$$\mu_2 = [-3, -5], \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



Expectation Maximization: a simple case

- A simple case:
  - We have unlabeled data $x^1, \cdots, x^m$
  - We know there are k clusters
  - We know all covariance matrices are the same

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

Start with an initial guess for $\mu_1, \ldots, \mu_k$, and $p(y=1), \ldots, p(y=k)$,
Repeat:

1. With current $\mu_1, \ldots, \mu_k$, we can easily compute probability that a point x^j belongs to class i:

$$p(y=i|x^j) \propto \exp\left(-\frac{1}{2\sigma^2}\left|x^j - \mu_i\right|^2\right) p(y=i)$$

E-step

Simply evaluate this, then normalize such that $\sum_{i=1,..,k} p(y=i|x^j) = 1$

2. With current estimate of $p(y=i|x^j)$ we can re-estimate the $\mu_1, \ldots, \mu_k$ and $p(y=1), \ldots, p(y=k)$

$$p(y=i) = \frac{\sum_{j=1}^{m} p(y=i|x^j)}{m} \qquad \mu_i = \frac{\sum_{j=1}^{m} p(y=i|x^j) x^j}{\sum_{j=1}^{m} p(y=i|x^j)}$$

M-step

Counting the points in each cluster Computing the center of each cluster

Similarity/Difference with Kmeans?

- E-step ≈ reassignment step
  - E-step assigns a point probabilistically
    - Clusters with closer center will get higher probability
    - How spread out the probability is depends on σ^2
    - In the extreme case, $\sigma^2 \to 0$, this will reduce to kmeans
  - K-Means reassign a point to a cluster deterministically
    - The closest cluster center gets 100%
- M-step ≈ recentering
  - M-step computes the center using weighted average

More General Cases

A simple case:

We have unlabeled data $x^1, \cdots, x^m$

We know there are k clusters

We know all covariance matrices are the same

Need to estimate $p_1, \ldots, p_k;\ \mu_1, \mu_2, \ldots, \mu_k$

General Case:

We have unlabeled data $x^1, \cdots, x^m$

We know there are k clusters

Each cluster has its own covariance matrix

Need to estimate $p_1, \ldots, p_k;\ \mu_1, \ldots, \mu_k; \Sigma_1, \ldots, \Sigma_k$

Expectation Maximization for the general case

Initialize, $p_1 p_2 \dots p_k, ; \mu_1 \dots \mu_k ; \Sigma_1 \dots \Sigma_k ,$

Iterate (till observe little change on parameters):

- E-step:

Compute for each point j and each cluster i

$$p(y = i | x^j) \propto \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (x^j - \mu_i)^T \Sigma_i^{-1} (x^j - \mu_i) \right) p(y = i)$$

Normalize such that $\sum_{i=1,..,k} p(y = i | x^j) = 1$

- M-step:

$$p_i = \frac{\sum_{j=1}^{m} p(y = i | x^j)}{m} \qquad \mu_i = \frac{\sum_{j=1}^{m} p(y = i | x^j) x^j}{\sum_{j=1}^{m} p(y=i|x^j)}$$

$$\Sigma_i = \frac{1}{n-1} \sum_{j=1}^{n} p(y = i | x^j)(x^j - \mu_i)(x^j - \mu_i)^T$$

You are not required to memorize the formulae, but its always good to know intuitively what they do.

Different Covariance Matrix Capture Different Shapes

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

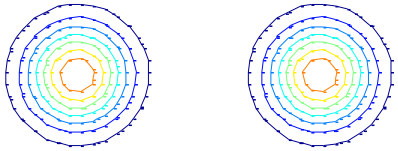
$\sigma^2 I$: a scaled identity matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

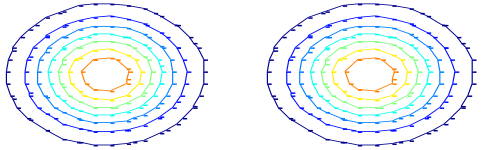
Diagonal matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

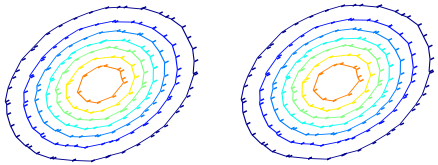
Full covariance matrix – most general



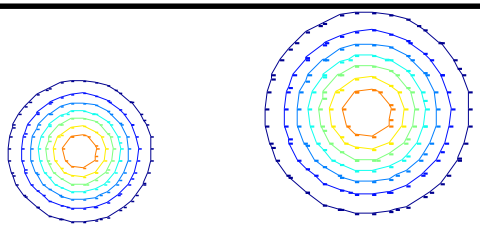
Different clusters with the same σ^2



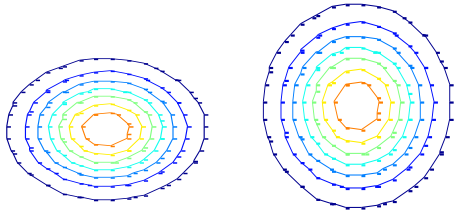
Different clusters with the same Σ



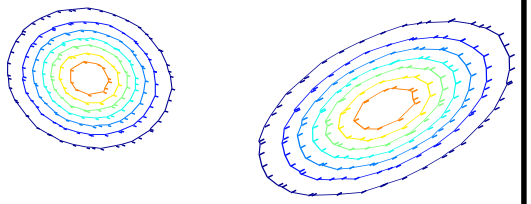
Different clusters with the same Σ



Different clusters with different σ^2 's

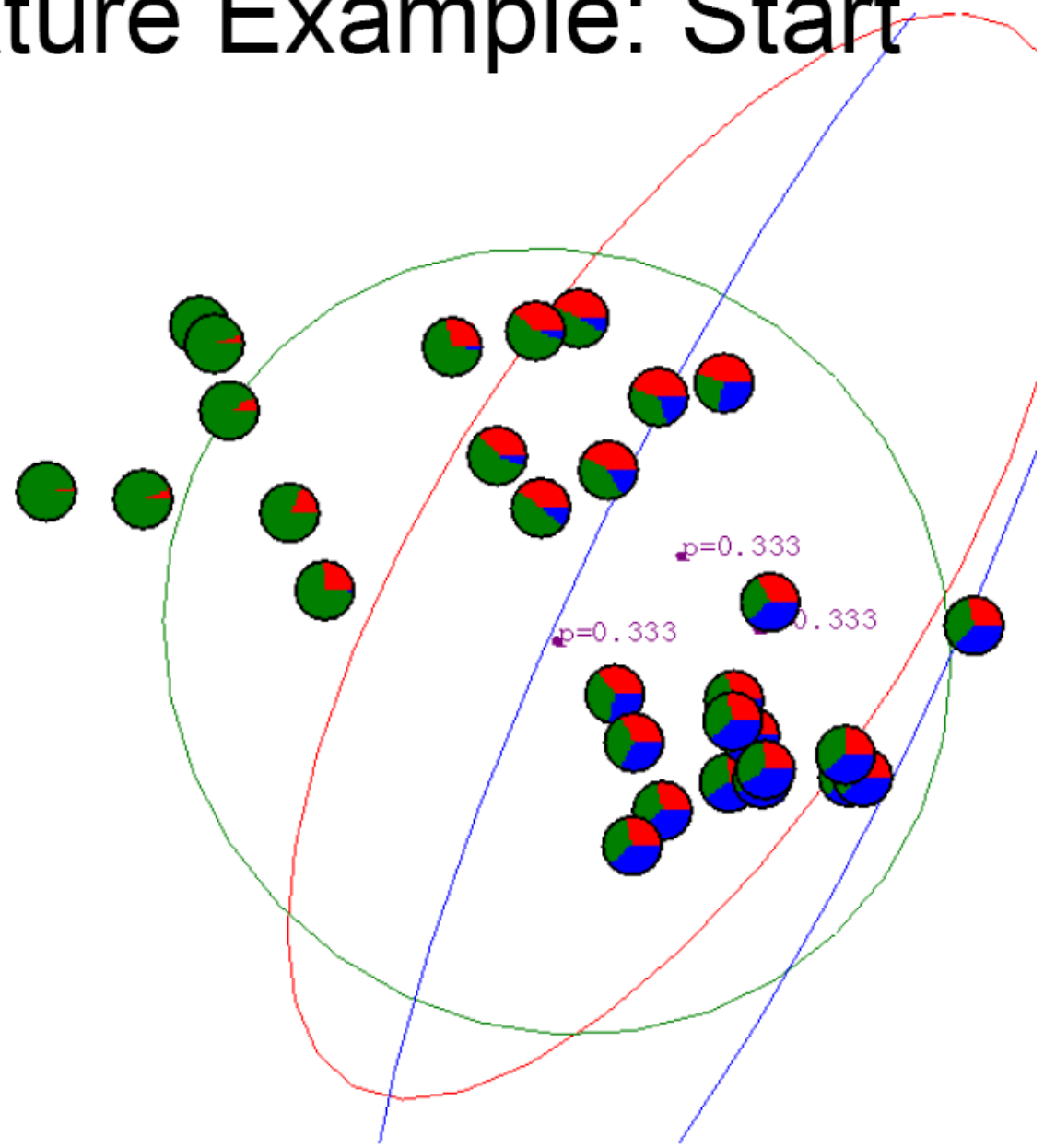


Different clusters with different Σ's

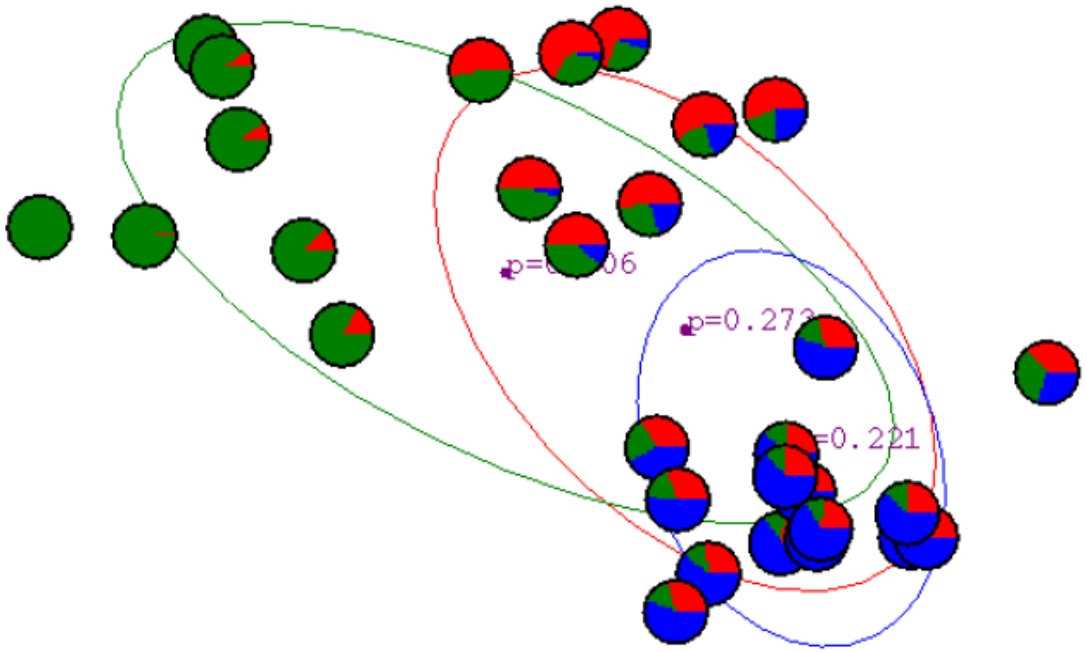


Different clusters with different Σ's

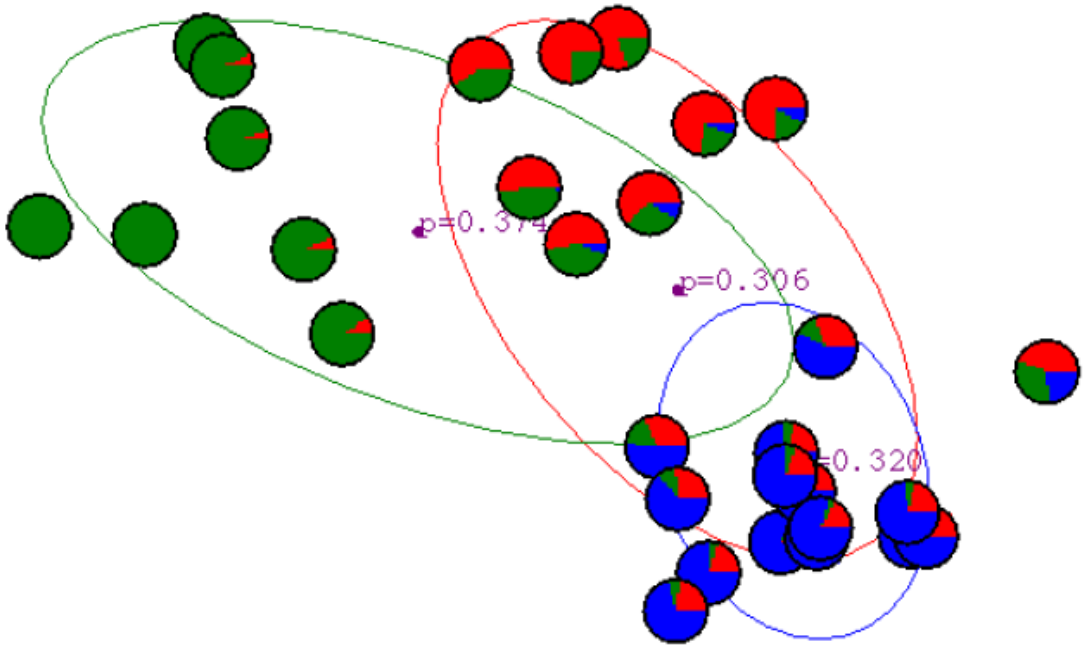
Gaussian Mixture Example: Start



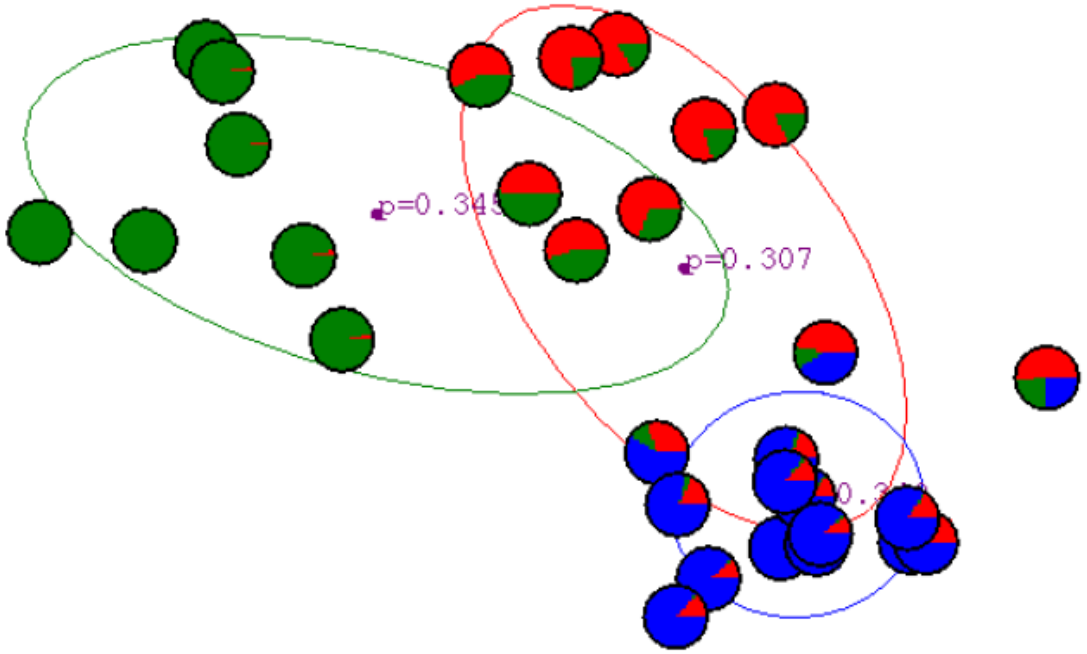
After first iteration



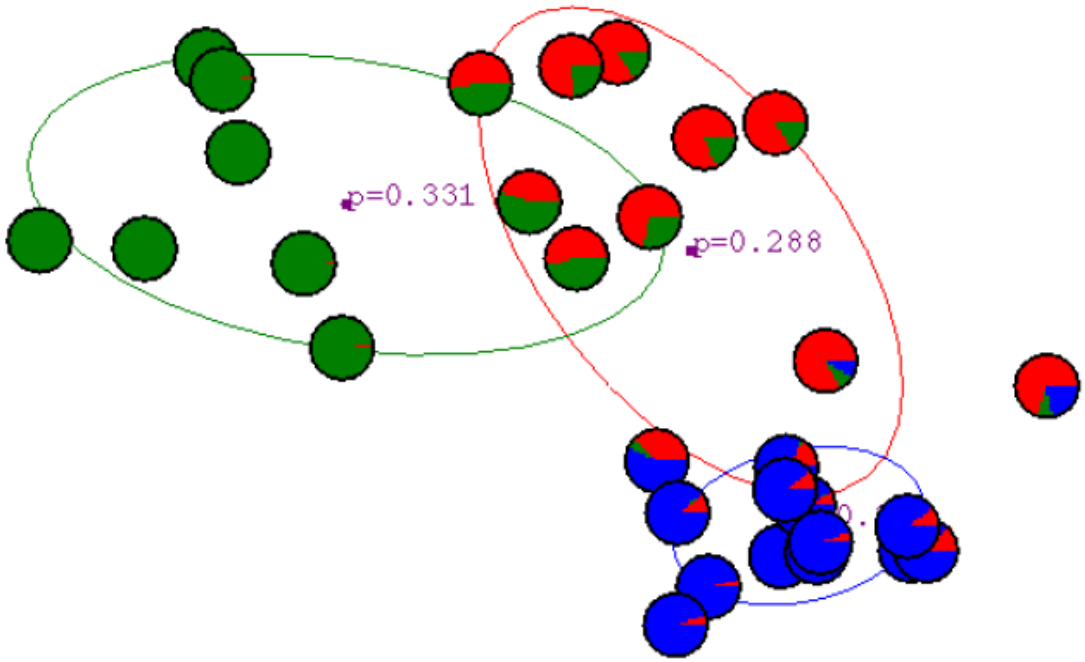
After 2nd iteration



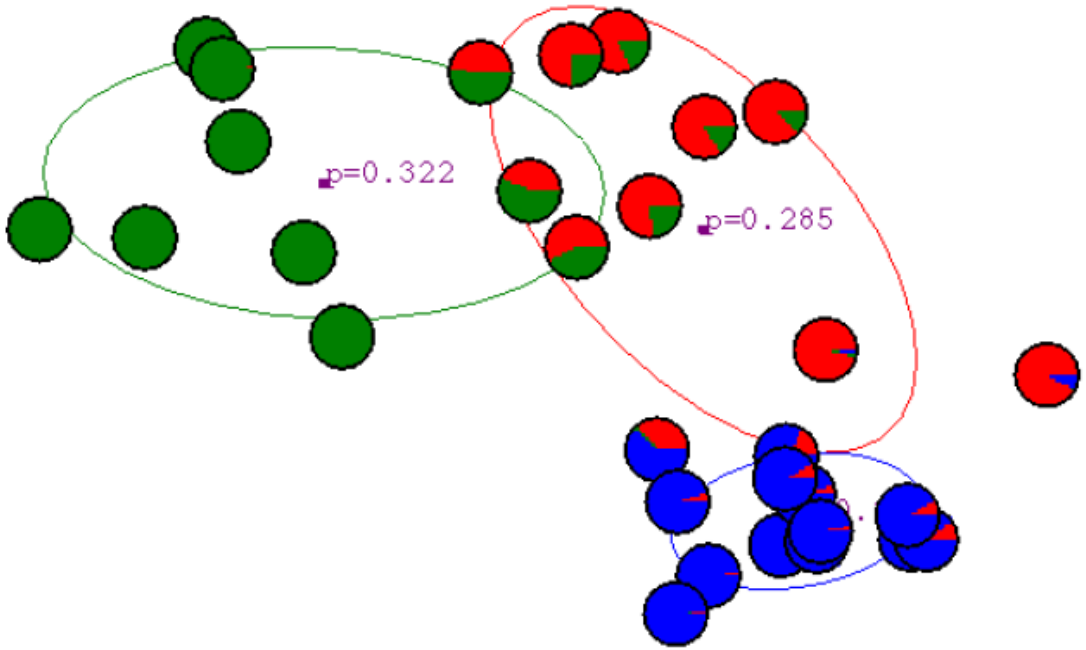
After 3rd iteration



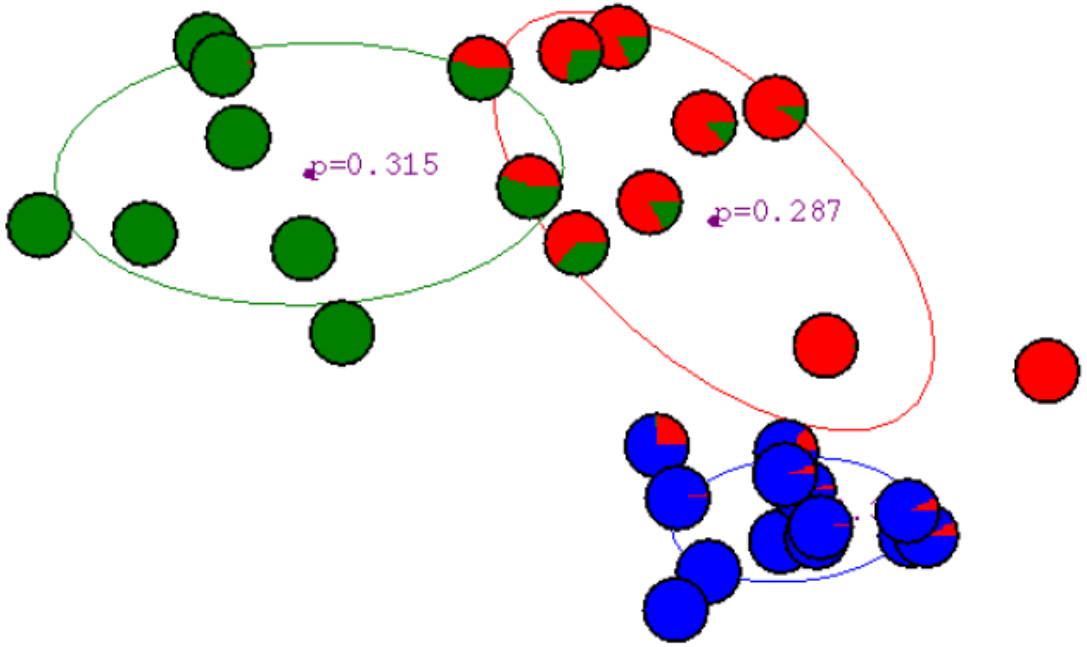
After 4th iteration



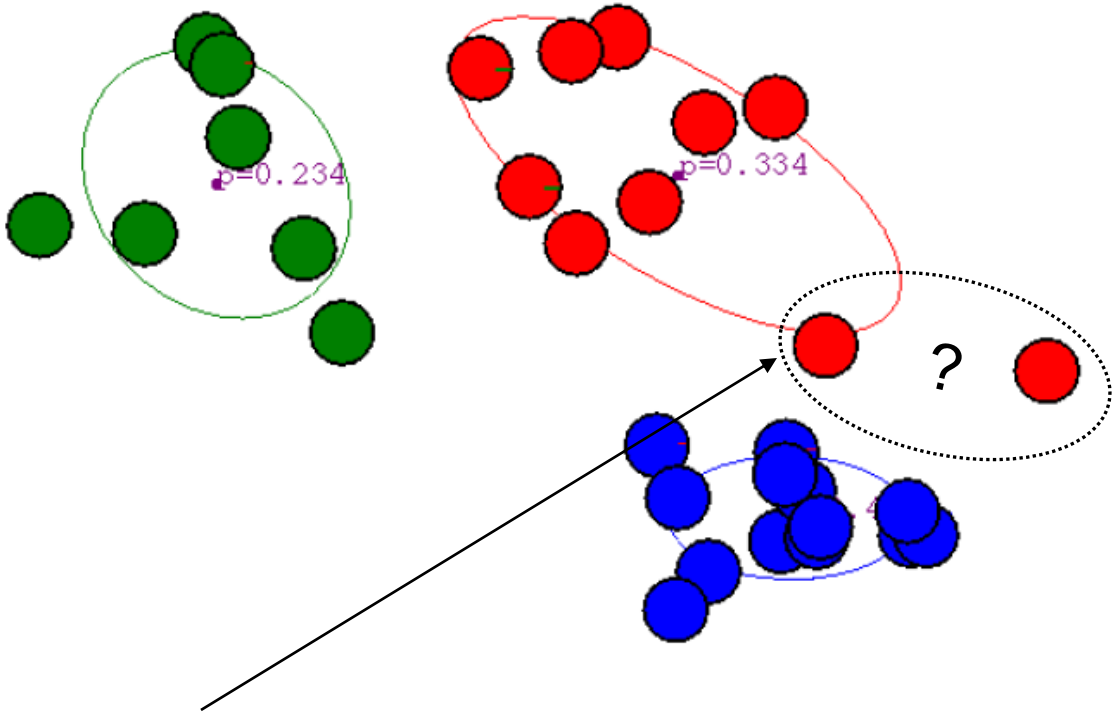
After 5th iteration



After 6th iteration



After 20th iteration



Q: Why are these two points red?

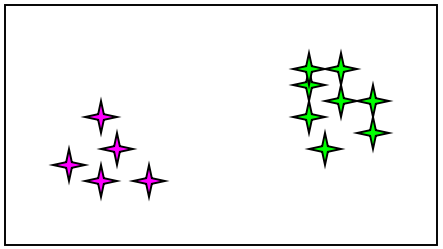
Behavior of EM

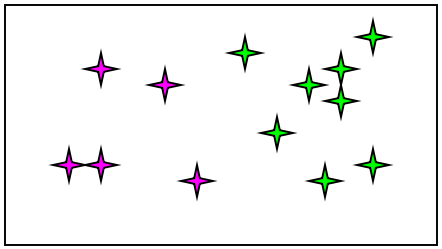
- It is guaranteed to converge
  - Convergence proof is based on the fact that $\log P(x|\theta)$ must increase or remain same between iterations (not obvious)
  - In practice it may converge slowly, one can stop early if the change in log-likelihood is smaller than a threshold
- It converges to a local optimum
  - Multiple restart is recommended
  - Choose the one that has the highest $\log P(x|\theta)$

Evaluating clustering

How to Evaluate Clustering?

- By user interpretation
  - does a document cluster seem to correspond to a specific topic?
- Internal criterion – a good clustering will produce high quality clusters:
  - high within-cluster similarity: $s_w = \sum_{i=1}^{k} \sum_{x,x' \in c_i} sim(x, x')$
  - low between-cluster similarity: $s_b = \sum_{x \in c_i, x' \in c_j, i \neq j} sim(x, x')$
  - E.g., s_w / s_b

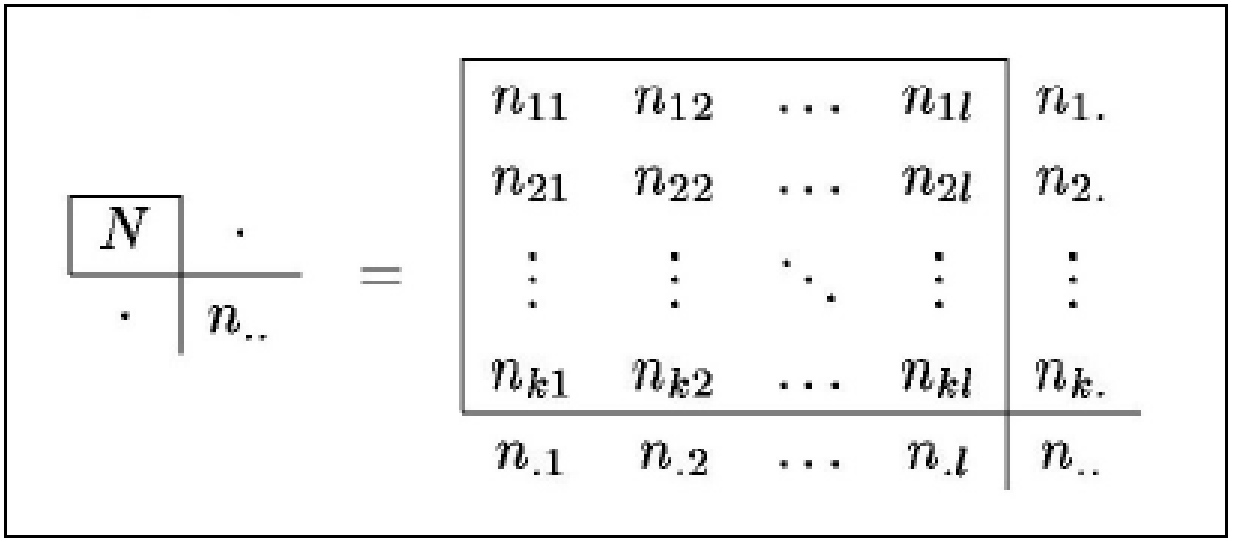




  - The measured quality of a clustering depends on both the object representation and the similarity measure used

External indexes

If true class labels (*ground truth*) are known, the validity of a clustering can be verified by comparing the class labels and clustering labels.



n_{ij} = number of objects in class *i* and cluster *j*

Rand Index and Normalized Rand Index

- Given partition (*P*) and ground truth (*G*), measure the number of vector pairs that are:
  - *a*: in the same class both in *P* and *G*.
  - *b*: in the same class in *P*, but different classes in *G*.
  - *c*: in different classes in *P,* but in the same class in *G*.
  - *d*: in different classes both in *P* and *G*.

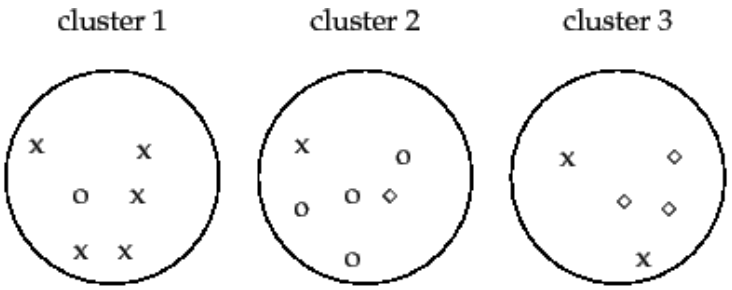
$$R = \frac{a + d}{a + b + c + d}$$

- Adjusted rand index: corrected-for-chance version of rand index
  - Compare to the expectation of the index assuming a random partition of the same cluster sizes

$$ARI = \frac{Index - ExpectedR}{MaxIndex - ExpectedR}$$

Purity and Normalized Mutual Information

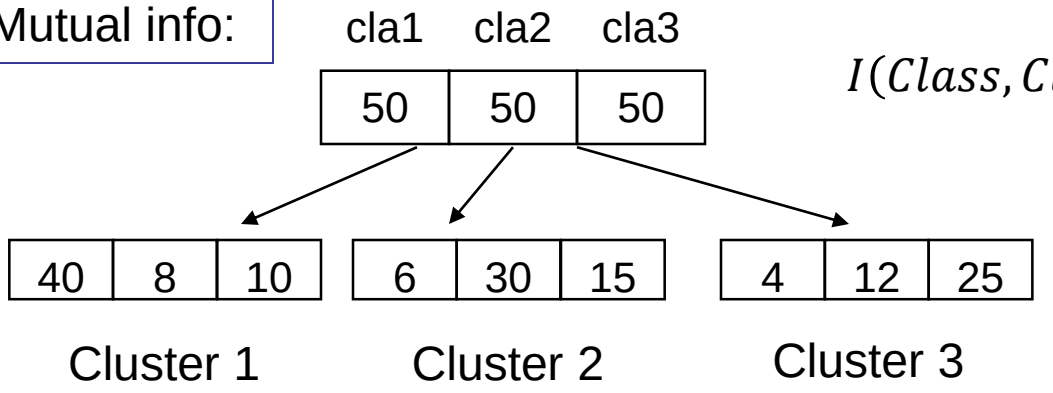
- Purity



► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and ⋄, 3 (cluster 3). Purity is $(1/17) \times (5+4+3) \approx 0.71$.

- Normalized Mutual Information

Mutual info:



$$I(Class, Clust) = H(Class) - H(Class|Clust)$$

$$NMI = \frac{2I(Class, Clust)}{H(Clust) + H(Class)}$$

Summary on clustering evaluation

- Internal evaluation
  - Within-cluster similarity, larger the better
  - Between-cluster similarity, smaller the better
  - Potential issue: Dependent on how similarity is measured
- External evaluation – compare against a ground truth clustering
  - Rand Index (very sensitive to the number of clusters)
  - Purity (very sensitive to the number of clusters)
  - Normalized Rand index (less sensitive to the number of clusters)
  - Normalized mutual information (less sensitive to the number of clusters)
  - Potential issue: there may not be only one way to cluster the data