# Expectation Maximization & Gaussian Mixture Models

ECE/CS 498 DS U/G

Lecture 7


Ravi K. Iyer

Electrical and Computer Engineering

University of Illinois

# Announcements

- MP1
  - Checkpoint 2 was due yesterday
  - Checkpoint 3 is due on Monday, Feb 18
  - Presentation on Friday, Feb 22 from 4pm – 6pm. Sign up for slots (first come first serve) when the poll is released on Wednesday
  - All group members must be present during presentation
- In class activity 2 on Bayesian Networks today
- Students are encouraged to answer/discuss questions on Piazza
  - Contribution will count towards class participation credit

# Looking for patterns and relationships in the data

- Clustering
  - Finding groupings in the data
  - It groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters
- Linear and non-linear regression
  - Finding relationship between different variables/features in the data
- Principal component analysis
  - Rotating the axes to simplify data visualization/description
  - Dimensionality reduction

# An illustration

- The data set has three natural groups of data points, i.e., 3 natural clusters
- “Similar” data points are more likely to belong to the same cluster compared to “dissimilar” data points

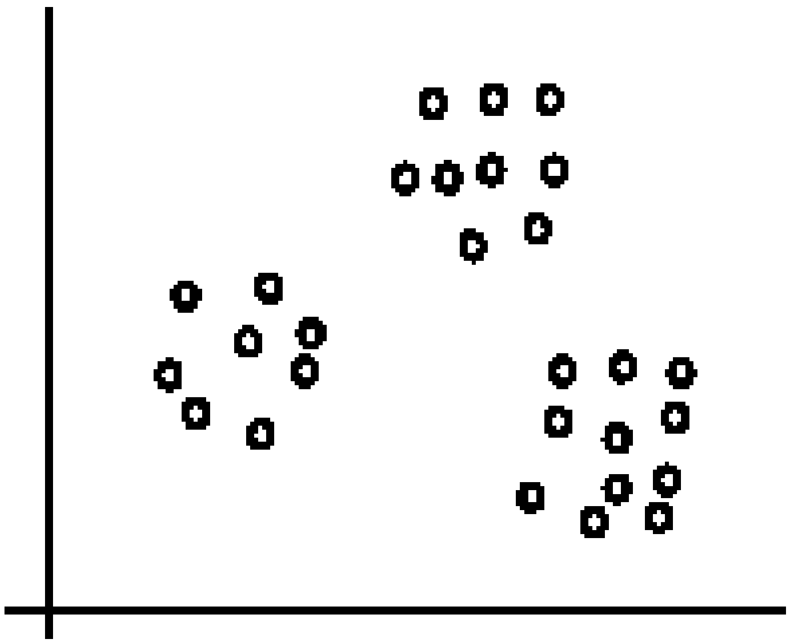
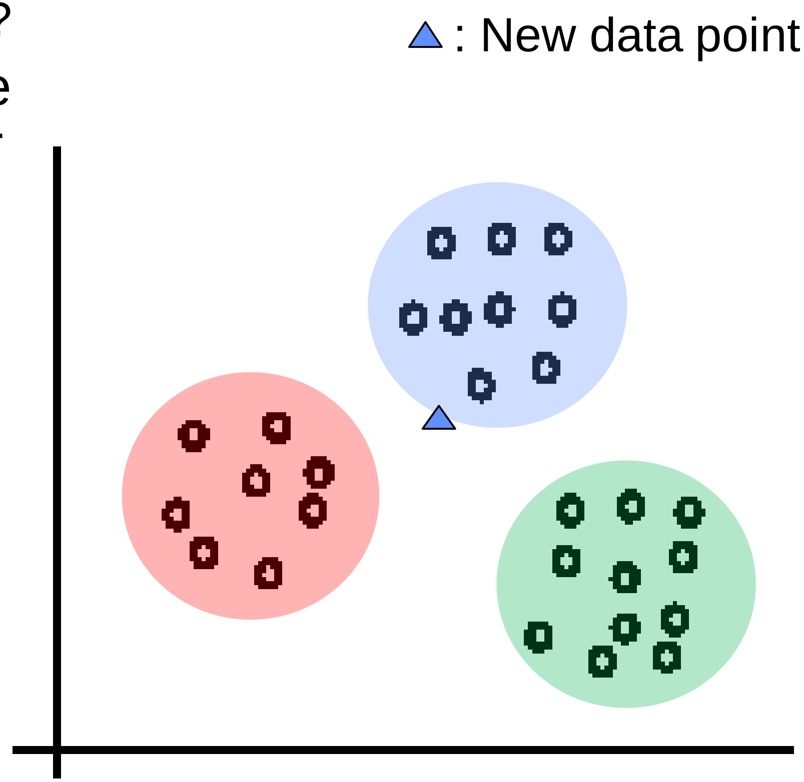


Image source: CS583, Bing Liu, UIC

# Aspects of Clustering: Algorithm

- Question:
  - How to find the clusters?
  - How to assign a point to a cluster?
- Want clusters of instances that are similar to each other but dissimilar to others

- Examples of methods:
  - Soft clustering
    - Gaussian Mixture Models
  - Hard clustering:
    - K-means clustering
    - Hierarchical clustering



# Aspects of Clustering: Distance function

- Question:
  - How similar is a point to a cluster or to the other points in a cluster?
- Need a similarity measure that measures how close a point is to a cluster or another point

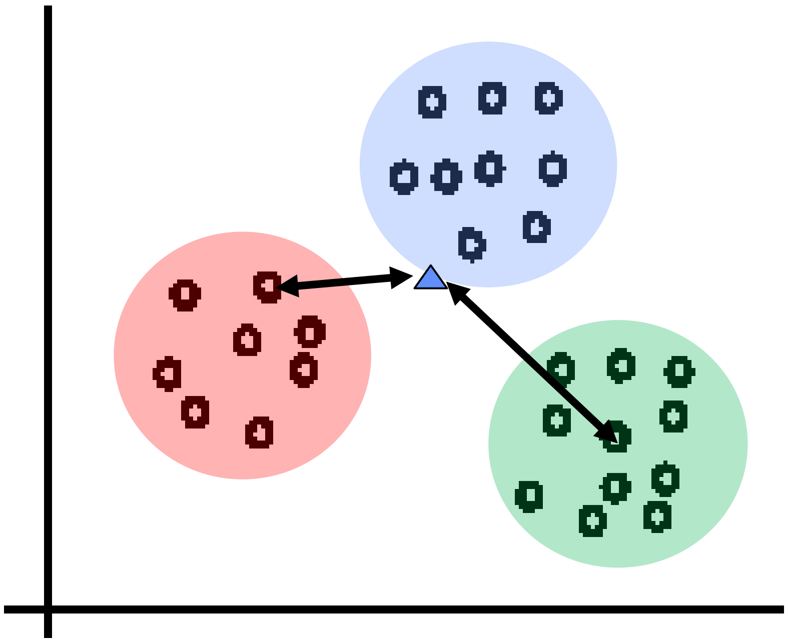
- Example of similarity measures when features are continuous
  Let $\boldsymbol{x_i}, \boldsymbol{x_j}$ be new points (features)
  - Euclidean distance measure (compact isolated clusters)

$$d(\boldsymbol{x_i}, \boldsymbol{x_j}) = \|\boldsymbol{x_i} - \boldsymbol{x_j}\|_2$$

  - The squared Mahalanobis distance alleviates problems with correlation

$$d(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\boldsymbol{x_i} - \boldsymbol{x_j})\Sigma^{-1}(\boldsymbol{x_i} - \boldsymbol{x_j})^T$$

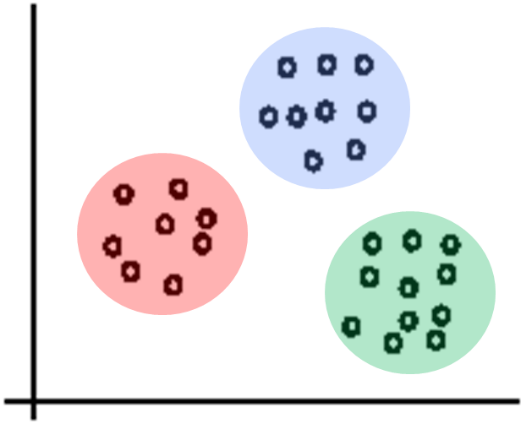
where $\Sigma$ is the covariance matrix



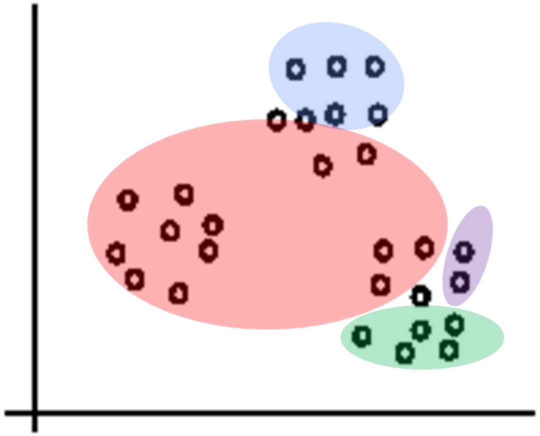
Example of Euclidean distance

# Aspects of Clustering: Cluster quality

- Clustering quality:
  - How "good" are the clusters?
- Examples of criteria:
  - Inter-clusters distance $\Rightarrow$ maximized
  - Intra-clusters distance $\Rightarrow$ minimized
- The quality of a clustering result depends on the algorithm, the distance function, and the data

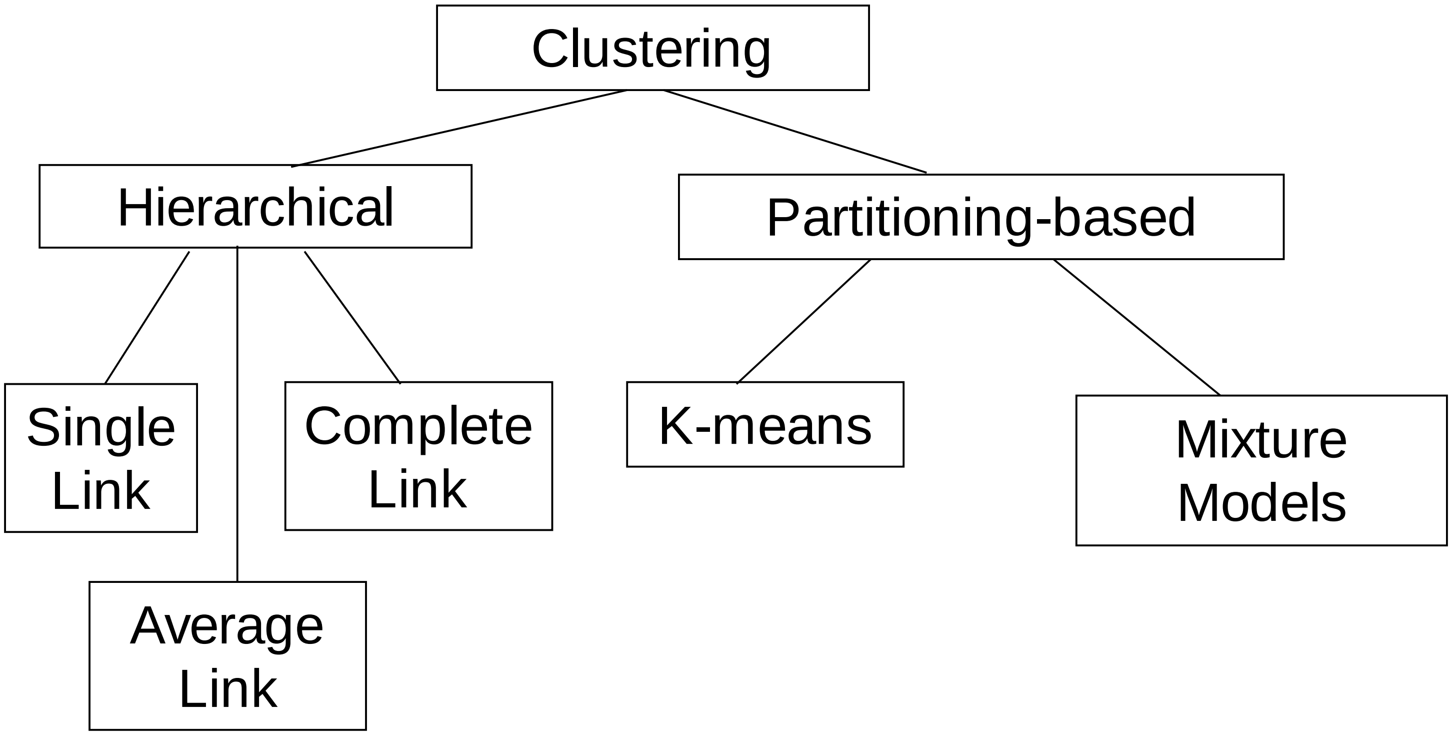


"Good" clusters



"Not-so-good" clusters

# Clustering Techniques



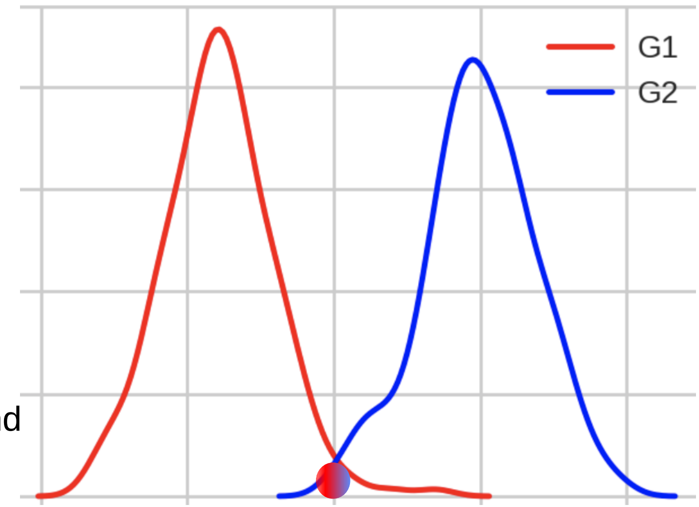
# Soft Clustering: Mixture Model: Clusters may Overlap

**Given:**

- Data points/observations: $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots$

**Model**:

- There is a set of *K* probability distributions
  - Each distribution represents a cluster
  - Each distribution is described by certain parameters
  - Clusters may overlap
  - Strengths of association between clusters and data instances
  - Discover the parameters of the distribution e.g. mean and variance
- Each data point is sampled from one of several distributions
  - $p(\boldsymbol{x_i}|b)$: Probability (density) that an instance $\boldsymbol{x_i}$ takes certain feature values given that it is from cluster $b$
  - $P(b|\boldsymbol{x_i})$: Probability that an instance belongs to cluster $b$ given that its features are $\boldsymbol{x_i}$




**Problem:**

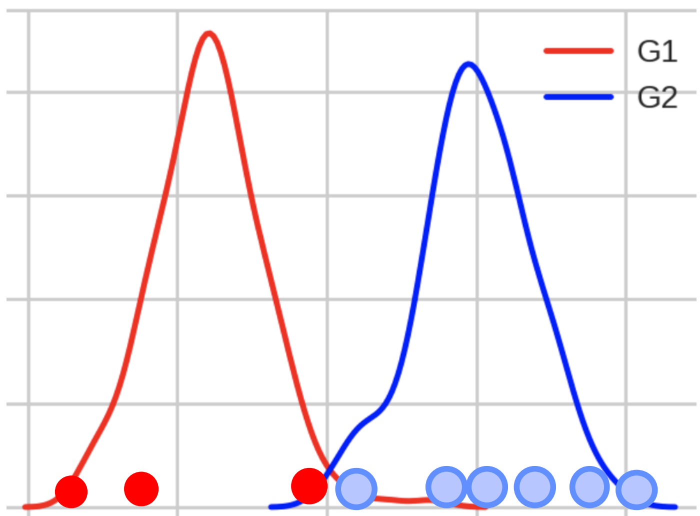
- Find parameters of the *K* distributions
- Find the posterior probabilities for each point

**Expectation Maximization**

- Automatically discover all the parameters for the K sources

# GMM Example: Find parameters

- Observations: $x_1, x_2, \ldots, x_N$
  - Each observation has 1 feature (1-dimension)
- Data is sampled from one of two Gaussian distributions (K=2)
  - Cluster $r$: $(\mu_r, \sigma_r^2)$
  - Cluster $b$: $(\mu_b, \sigma_b^2)$




- **Estimation**: If source (cluster) of each observation is known, it is trivial to estimate $(\mu_r, \sigma_r^2)$ and $(\mu_b, \sigma_b^2)$

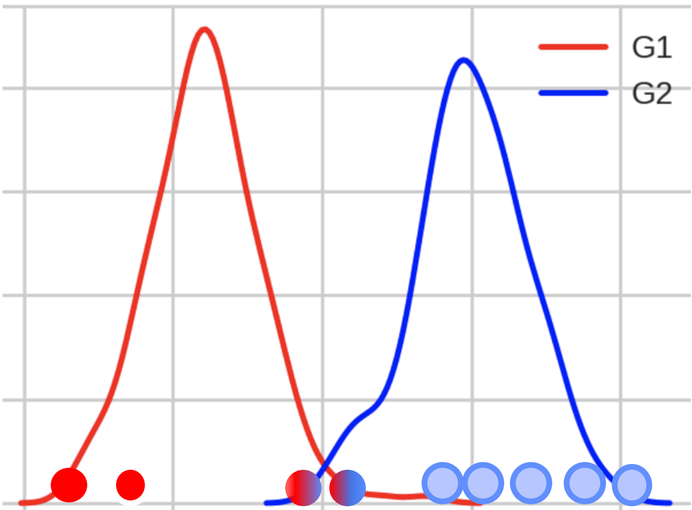
$$\mu_r = \frac{\sum_{i=1}^{N} x_i \mathbb{I}\{x_i \sim r\}}{\sum_{i=1}^{N} \mathbb{I}\{x_i \sim r\}} \qquad \sigma_r^2 = \frac{\sum_{i=1}^{N} (x_i \mathbb{I}\{x_i \sim r\} - \mu_a)^2}{\sum_{i=1}^{N} \mathbb{I}\{x_i \sim r\}}$$

$$\mu_b = \frac{\sum_{i=1}^{N} x_i \mathbb{I}\{x_i \sim b\}}{\sum_{i=1}^{N} \mathbb{I}\{x_i \sim b\}} \qquad \sigma_b^2 = \frac{\sum_{i=1}^{N} (x_i \mathbb{I}\{x_i \sim b\} - \mu_b)^2}{\sum_{i=1}^{N} \mathbb{I}\{x_i \sim b\}}$$

where $\mathbb{I}\{x_i \sim r\} = 1$ if $x_i$ was sampled from cluster $r$ and 0 otherwise.

# GMM Example: Find posterior

- Observations: $x_1, x_2, \dots, x_N$
  - Each observation has 1 feature (1-dimension)
- Data is sampled from one of two Gaussian distributions (K=2)
  - Cluster $a$: $(\mu_a, \sigma_a^2)$
  - Cluster $b$: $(\mu_b, \sigma_b^2)$




- If the distribution and its parameters is known, estimate where the point is likely to come from using Bayes rule

$$P(b|x_i) = \frac{p(x_i|b)P(b)}{p(x_i|b)P(b) + p(x_i|r)P(r)}$$

Posterior probability of distribution $b$ given sample $x_i$

$$p(x_i|b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$

Probability density of observing $x_i$ when sampled from distribution $b$

# Expectation Maximization

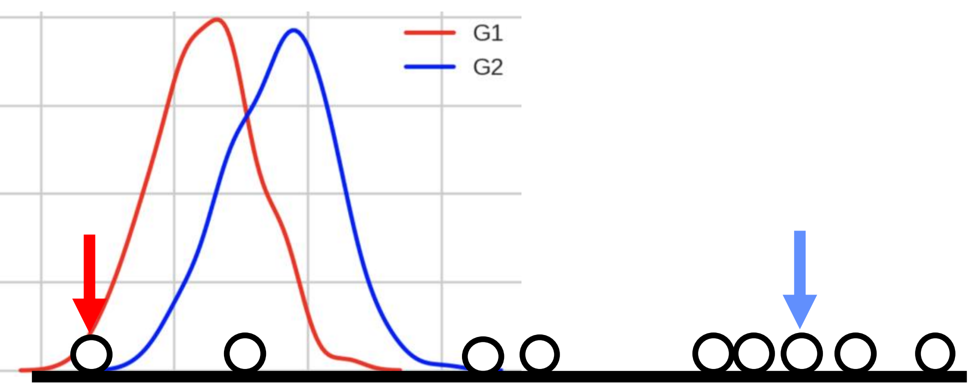
- **What if neither the source nor the distribution parameters are known?**

- **Chicken and Egg problem**
  - Need $(\mu_b, \sigma_b^2)$ and $(\mu_r, \sigma_r^2)$ to guess source of points
  - Need to know source to estimate $(\mu_b, \sigma_b^2)$ and $(\mu_r, \sigma_r^2)$
  - Use **Expectation Maximization (EM)** algorithm

- **EM Algorithm**
  - Start with **two randomly placed Gaussians** $(\mu_b, \sigma_b^2)$ and $(\mu_r, \sigma_r^2)$
  - For each $x_i$, calculate $P(b|x_i)$ and $P(r|x_i) = 1 - P(b|x_i)$
  - **Remember it does not assign the point but says here is the probability that it came from the red or from the blue**
  - Adjust $(\mu_b, \sigma_b^2)$ and $(\mu_r, \sigma_r^2)$ to fit points assigned to them

# GMM Example: EM in action

- Start with **two randomly placed Gaussians** $(\mu_b, \sigma_b^2)$ and $(\mu_r, \sigma_r^2)$
- **Expectation step (E)**: Assign posterior probabilities to each sample $x_i$




$P(r|x) > P(b|x)$ $\qquad$ $P(b|x) > P(r|x)$

- Let $b_i$ be the posterior probability of sample $x_i$ of belonging to cluster $b$

$$b_i = P(b|x_i) = \frac{p(x_i|b)P(b)}{p(x_i|b)P(b) + p(x_i|r)P(r)}$$

$$p(x_i|b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right)$$

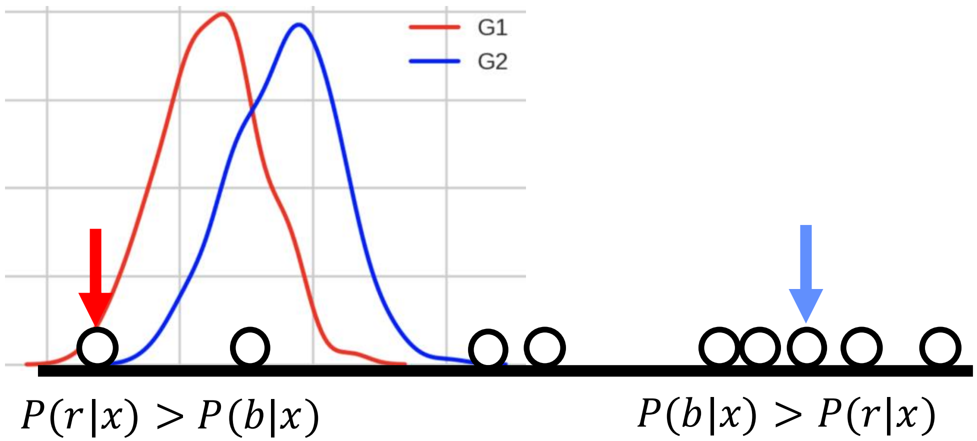
Probability density of observing $x_i$ when sampled from distribution $b$

- Similarly, let $r_i$ be the posterior probability of sample $x_i$ belonging to cluster $r$

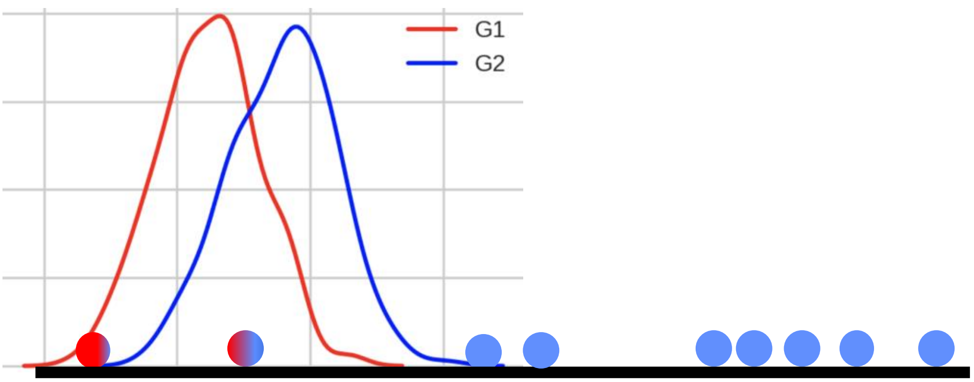
$$r_i = 1 - b_i$$

# GMM Example: EM in action

Before assigning posterior probabilities $b_i$ and $r_i$




After assigning posterior probabilities $b_i$ and $r_i$



# GMM Example: EM in action

- **Maximization step (M)**: Update the distribution parameters (re-estimation)
- Take weight average of the samples
  - Weight is the posterior probability of that sample
- Similar to previous estimation with $\mathbb{I}\{x_i \sim b\}$ replaced by $P(b|x_i)$
  - $P(b|x_i)$ gives the likely is it that $x_i$ belong to $b$
  - Therefore, $x_i$'s contribution in re-estimating the parameters for $b$ is $P(b|x_i)$

$$\mu_b = \frac{b_1 x_i + b_2 x_2 + \cdots + b_N x_N}{b_1 + b_2 + \cdots + b_N} = \frac{\sum_{i=1}^{N} b_i x_i}{\sum_{i=1}^{N} b_i}$$

$$\sigma_b^2 = \frac{b_1 (x_1 - \mu_b)^2 + b_2 (x_2 - \mu_b)^2 + \cdots + b_N (x_N - \mu_b)^2}{b_1 + b_2 + \cdots + b_N}$$

$$= \frac{\sum_{i=1}^{N} b_i (x_i - \mu_b)^2}{\sum_{i=1}^{N} b_i}$$

$$P(b) = \frac{b_1 + b_2 + \cdots + b_N}{N} = \frac{\sum_{i=1}^{N} b_i}{N}$$

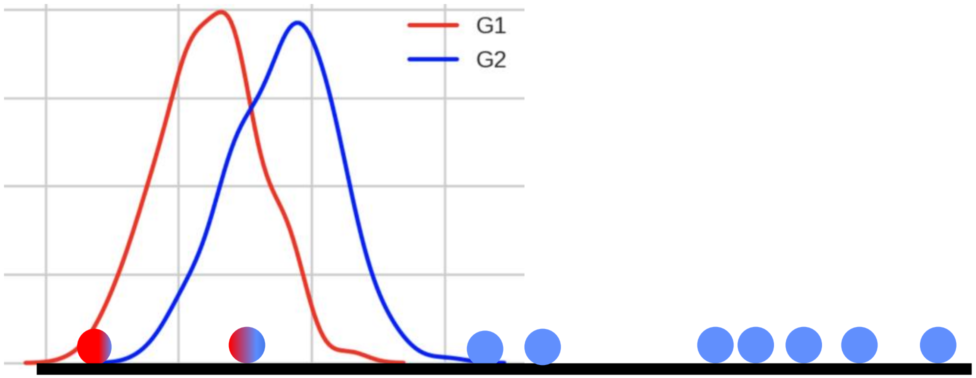
$$\mu_r = \frac{\sum_{i=1}^{N} r_i x_i}{\sum_{i=1}^{N} r_i}$$

$$\sigma_r^2 = \frac{\sum_{i=1}^{N} r (x_i - \mu_r)^2}{\sum_{i=1}^{N} r_i}$$

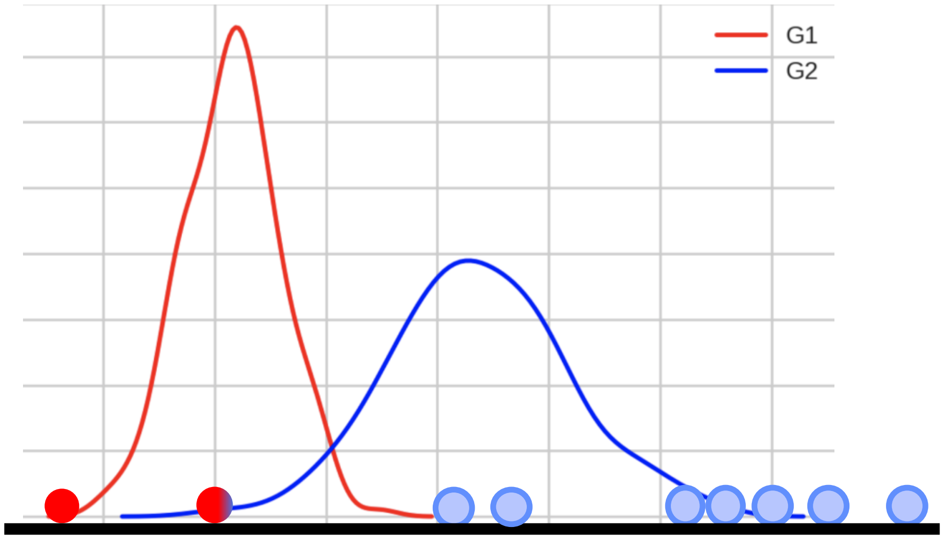
$$P(r) = \frac{\sum_{i=1}^{N} r_i}{N}$$

# GMM Example: EM in action

Distributions before updating their parameters

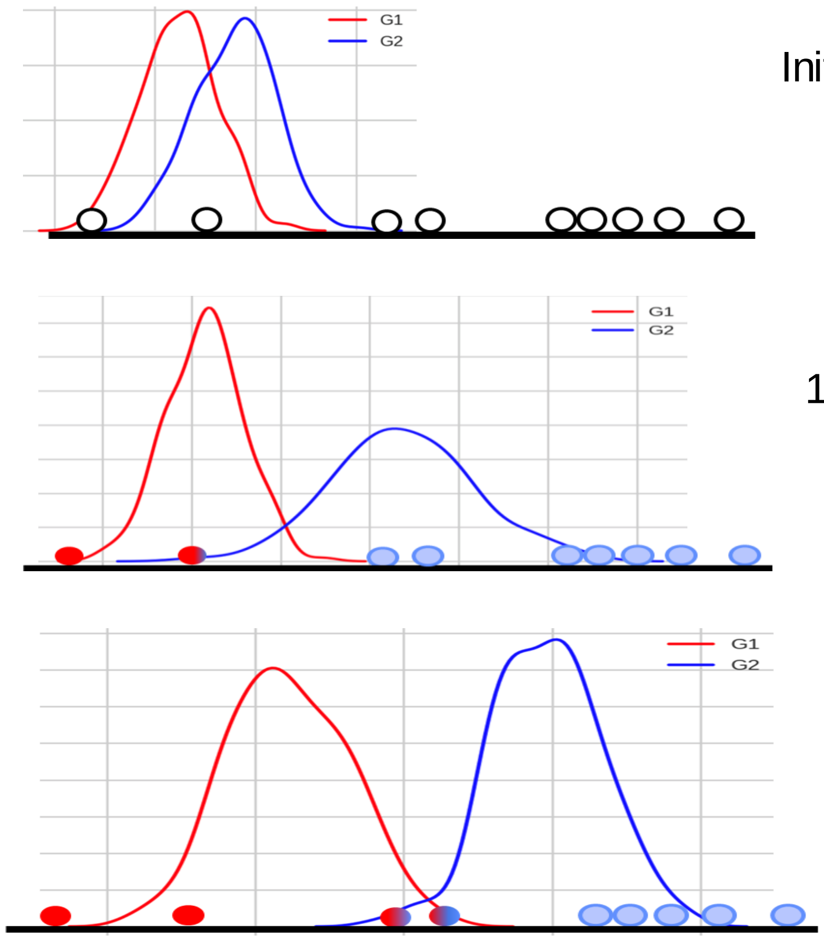



Distributions after updating their parameters using the posteriors



# GMM Example: EM in action

- Repeat the E and M steps iteratively till convergence
- Convergence: When M step gives the same parameters that were used in E



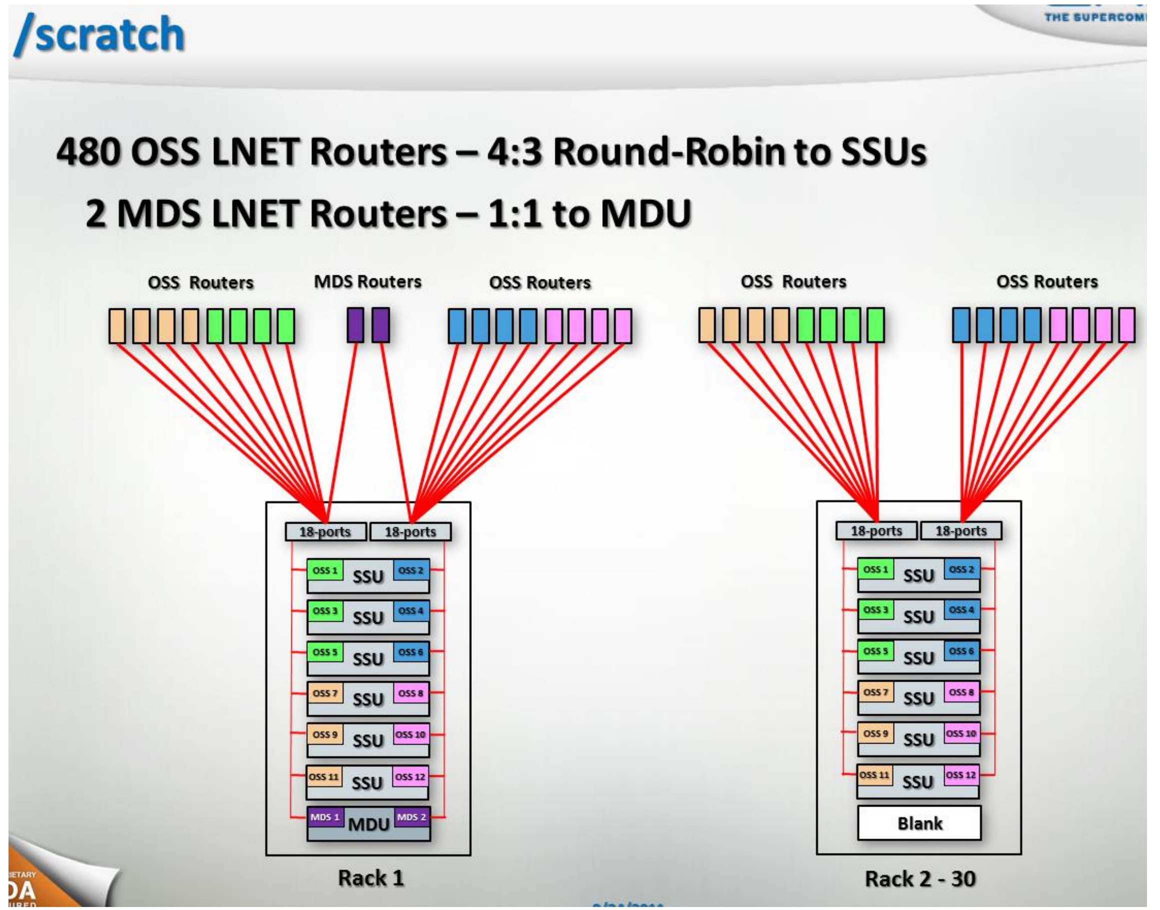

Initialization

1 iteration

Convergence ($n$ iteration)

# GMM Example : diagnosing the failing storage servers

- Blue Waters has 360 storage server for /scratch (22 PB usable)
- 30 racks * 6 SSU's per/rack = 180 SSU's
- 2 OSS/SSU, 180 *2 = 360 OSS
- 1440 * 20-TByte drives (7200 RPM NL-SAS) = 29 PBytes raw



# GMM Example : diagnosing the Failing storage servers

- Detecting anomalously behaving (i.e., unhealthy) storage servers
- Blue Waters has 362 storage server for /scratch (22 PB)
- Features
  - Time between Errors
  - Server Load