CLUSTERING

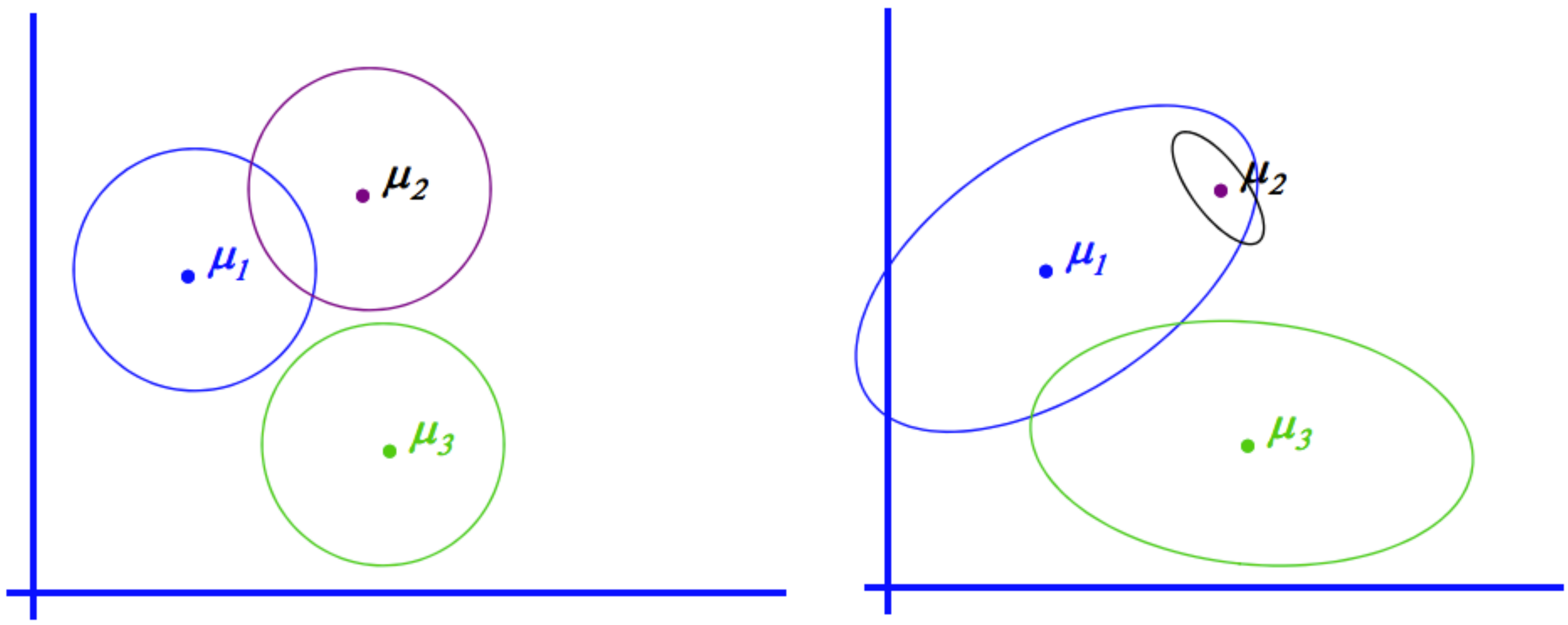
Oct 30th & 31st

Daegun Won

Outline

- Clustering review
  - K-means
  - GMM
  - Hierarchical clustering
- Examples

K-means & GMM

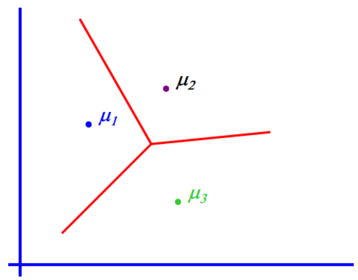


K-means

- (Randomly) initialize k centers
  - $\mu^{(0)} = \mu_1^{(0)}, \mu_2^{(0)}, \ldots, \mu_k^{(0)}$
- Assign each point *j* to nearest center
  - $C^{(t)}(j) = \text{argmin}_i \left\| \mu_i^{(t)} - x_j \right\|^2$
- Reposition the centers
  - $\mu_i^{(t+1)} = \text{argmin}_\mu \sum_{j:C(j)=i} \left\| \mu - x_j \right\|^2$
- Repeat until none of the assignments change

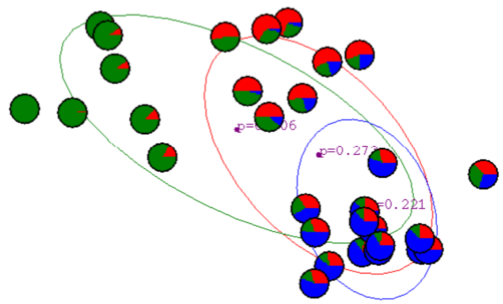
K-means (2)

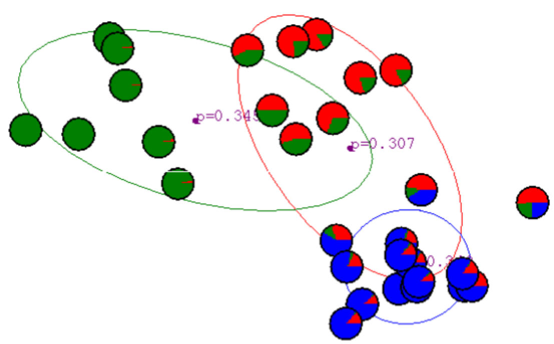
- Linear decision boundary
  - Voronoi diagram
  - Clusters may not be linearly separable
- Hard assignments
  - Clusters may overlap
- Same diagonal covariance matrix
  - Some clusters may be wider than others



GMM

- (Randomly) initialize
- (E-step) Do “soft” assignment of each point
  - $$P\left(y = i \middle| x_j, \theta^{(t)}\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|x_j - \mu_i^{(t)}\right\|^2\right) P^{(t)}(y = i)$$





GMM(2)

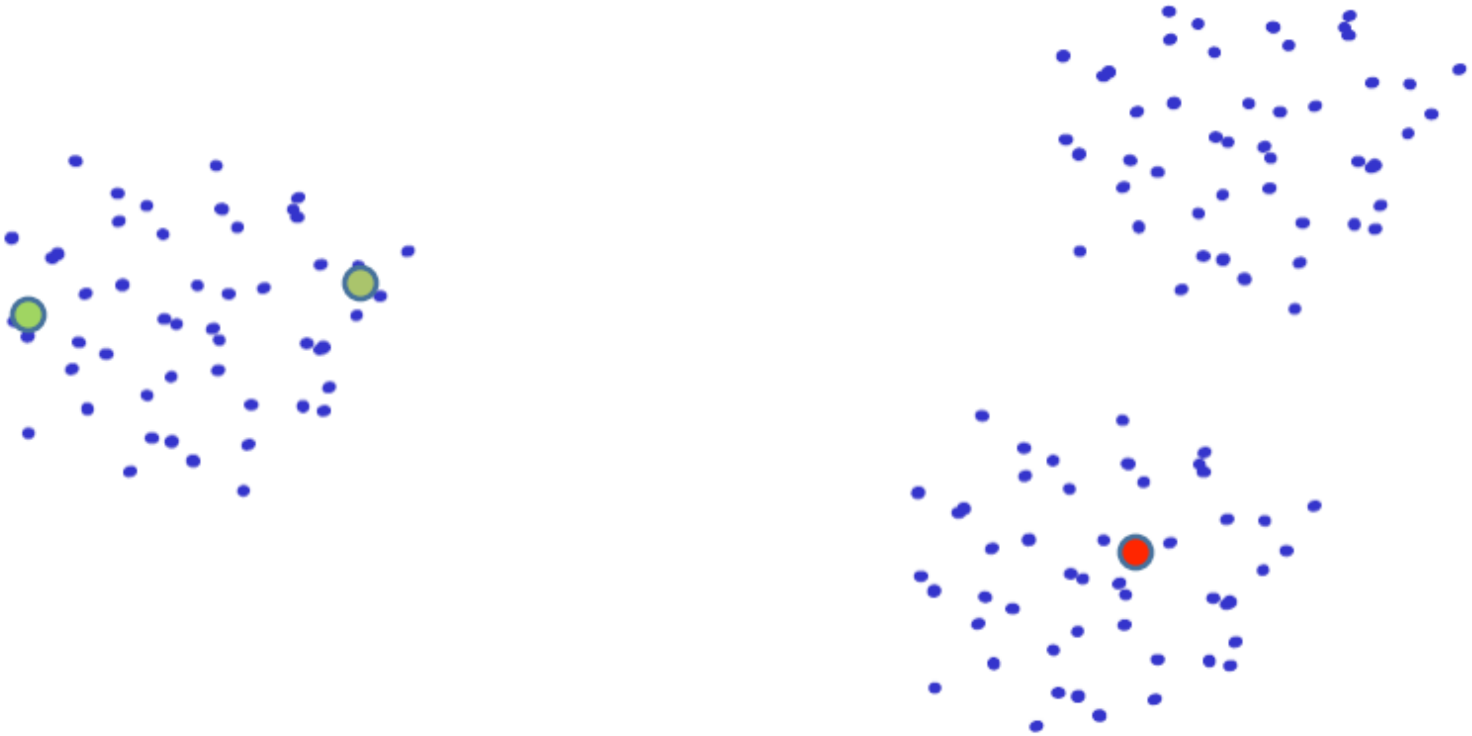
- (M-step) Compute MLEs
  - $\mu_i^{(t+1)} = \frac{\sum_j P(y=i|x_j,\theta^{(t)})x_j}{\sum_j P(y=i|x_j,\theta^{(t)})}$
  - $\Sigma_i^{(t+1)} = \frac{\sum_j P(y=i|x_j,\theta^{(t)})(x_j-\mu_i^{(t+1)})(x_j-\mu_i^{(t+1)})^T}{\sum_j P(y=i|x_j,\theta^{(t)})}$
  - $P^{(t+1)}(y = i) = \frac{\sum_j P(y=i|x_j,\theta^{(t)})}{\#\text{data points}}$

Hierarchical Clustering

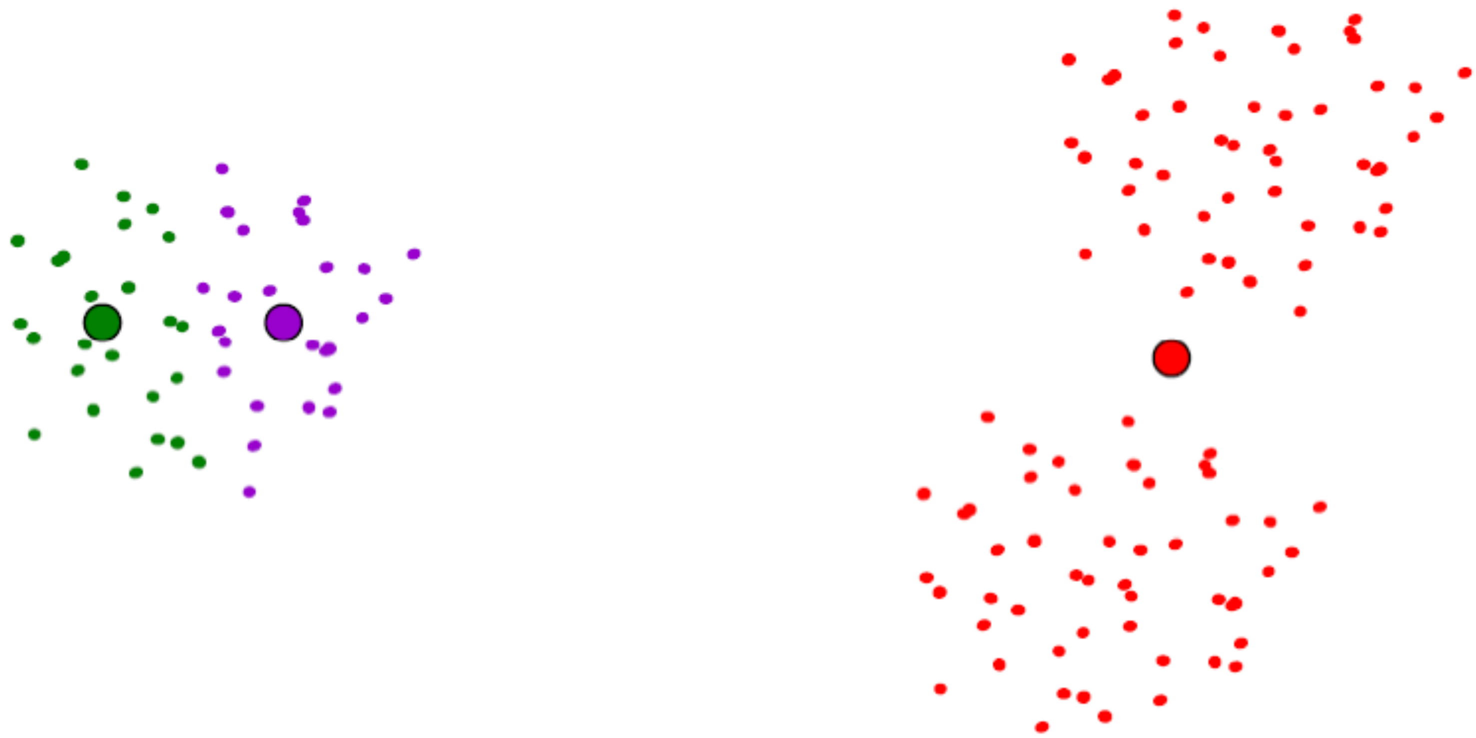
- Bottom-up (agglomerative) clustering
  - Find a pair of clusters to merge into a new cluster that has the shortest cluster distance
  - Cluster distance metrics
    - Single link: distance of the two closest members in each cluster
    - Complete link: distance of two farthest members
    - Average: average distance of all pairs

LET'S SEE SOME EXAMPLES!

K-means – seed choice?



Oops!



K-means

- What’s the effect on the means found by k-means (as opposed to the true means) of overlapping clusters?

K-means

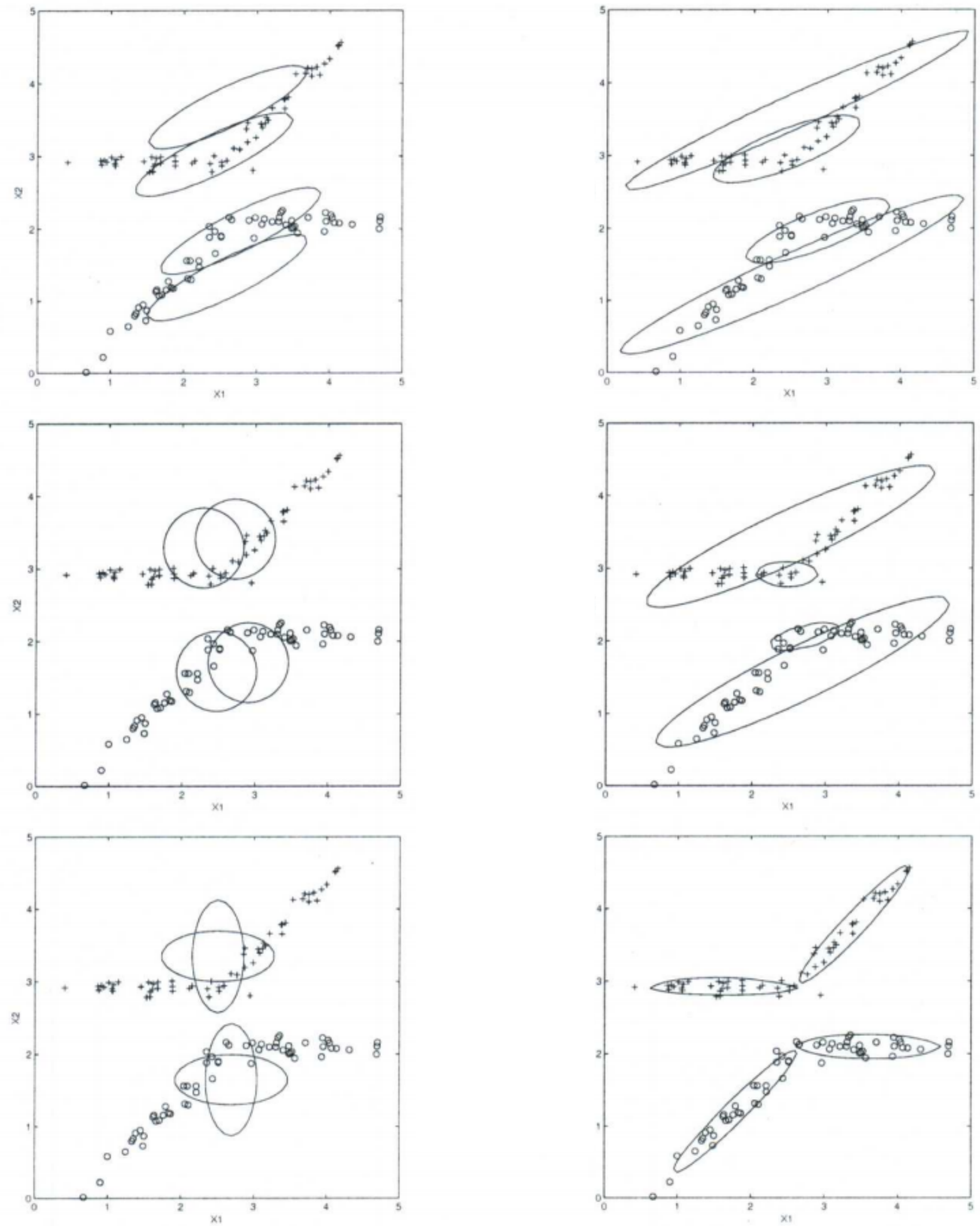
- What's the effect on the means found by k-means (as opposed to the true means) of overlapping clusters?
  - The means found by k-means will be further apart

GMM

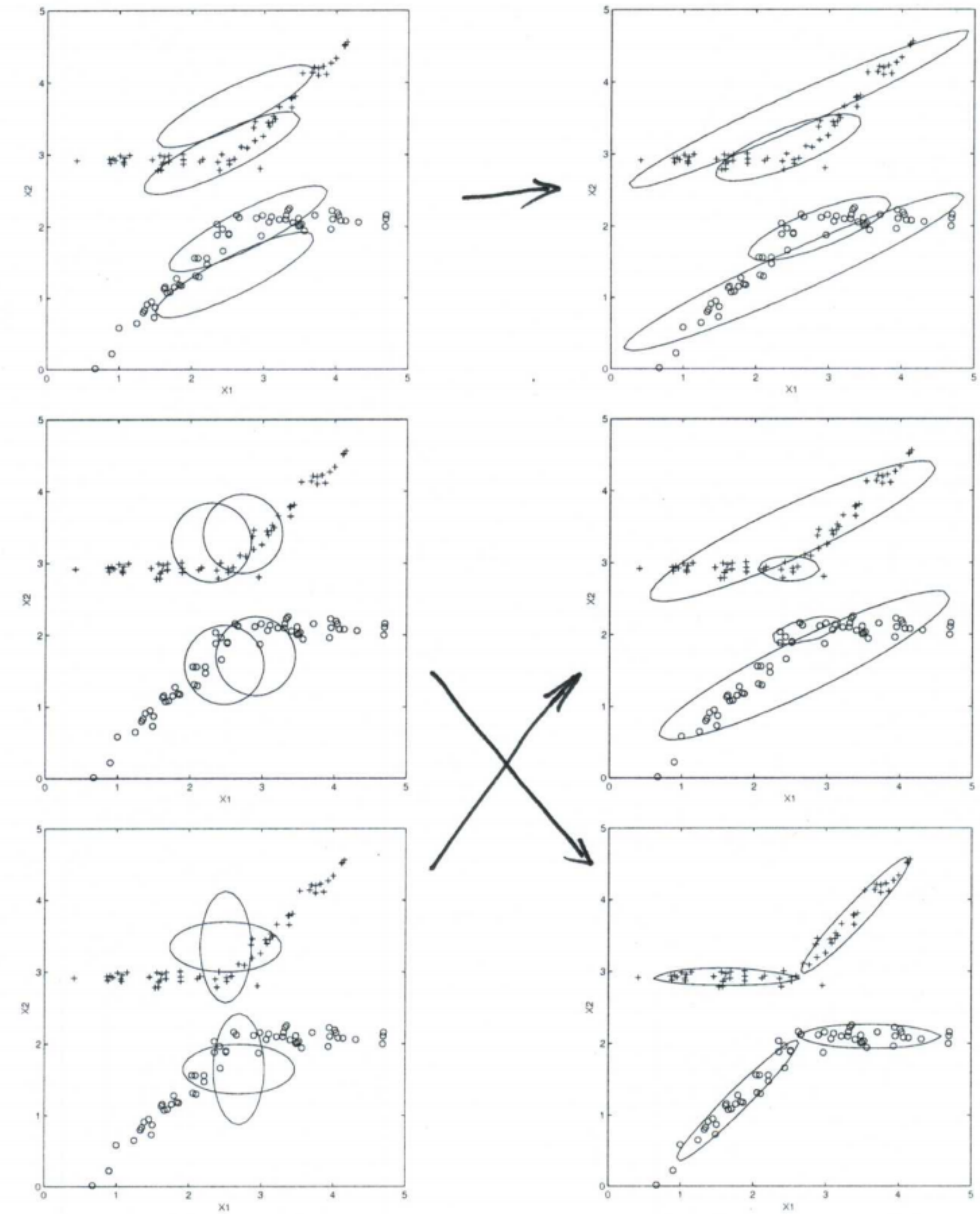
- Suppose a GMM has two components
$$0.5N(\mu_1, 1) + 0.5N(\mu_2, 1)$$
- The observed data are $x_1 = 2, x_2 = 0.5$ and the current estimates of μ_1 and μ_2 are 2 and 1. Compute the component memberships of this observed data for the next E-step.
  - (normal densities for standard normal distribution at 0, 0.5, 1, 1.5, 2 are 0.4, 0.35, 0.24, 0.13, 0.05)

- $P(z_1|x_1)$
  - $\frac{P(x_1|z_1)P(z_1)}{P(x_1|z_1)P(z_1)+P(x_1|z_2)P(z_2)} = \frac{0.4\times0.5}{0.4\times0.5+0.24\times0.5} = \frac{5}{8}$
- $P(z_2|x_1)$
  - $1 - P(z_1|x_1) = \frac{3}{8}$
- $P(z_1|x_2)$
  - $\frac{P(x_2|z_1)P(z_1)}{P(x_2|z_1)P(z_1)+P(x_2|z_2)P(z_2)}$

GMM



(a) Initialization (b) After first iteration



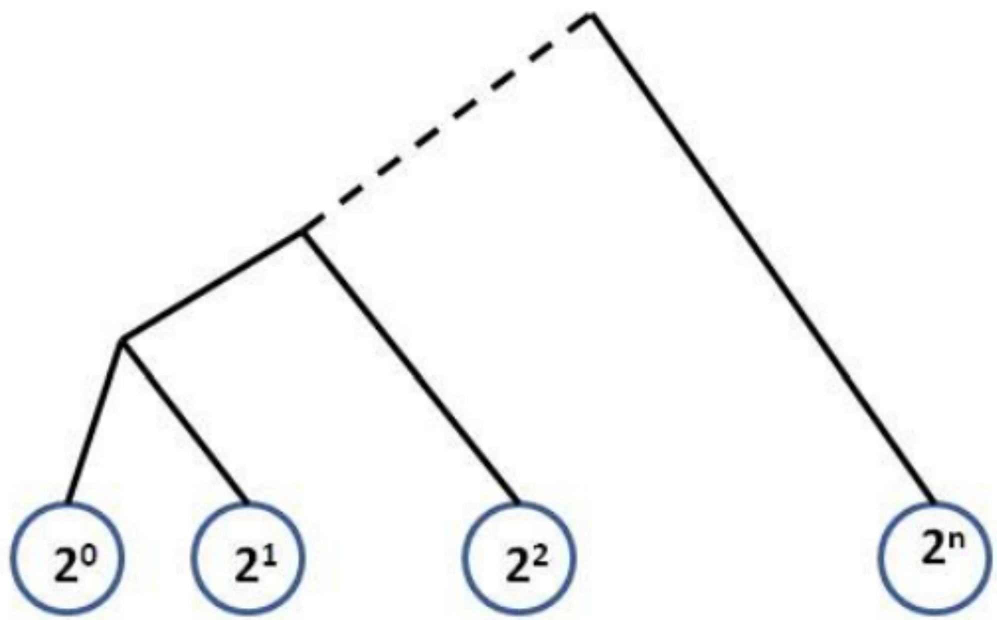
(a) Initialization (b) After first iteration

Hierarchical Clustering

- Assume we are trying to cluster n+1 points (2^0, 2^1, … 2^n) on one-dimensional space using hierarchical clustering.

  Suppose we use Euclidean distance and draw a sketch of the clustering tree we would get for
  - Single link
  - Complete link
  - Average link
- Assume we're looking at instances/clusters from left(small) to right(big) along the axis.

- For all linkage methods,



Hierarchical Clustering(2)

- Suppose we use the following distance function instead of Euclidean distance:

  $d(A,B) = |\log A - \log B|$

  Will the clustering result change from the previous question?
  - Single link
  - Complete link
  - Average link

- Single link won't change
- Complete link & average link will result in a perfect binary tree shape.