Quiz 2 Review

Lecture 4

Quiz

- Format: Blackboard, so remember…
  - Bring your laptop charged
  - Install LockDown Browser
- No notes/programs/web/calculator/etc
  - Scratch paper will be supplied
  - Bring a writing utensil
- Content: all of clustering

Question

- Of the four algorithms we covered (K-Means, agglomerative, DBSCAN, GMMs), which (if any) could have led to the following clustering (to 3 decimal places of precision)

	c1	c2	c3
p1	0	0	1
p2	0	1	0
p3	1	0	0
p4	0	1	0

Answer

- K-Means
- DBSCAN
- GMMs

Question

- Of the four algorithms we covered (K-Means, agglomerative, DBSCAN, GMMs), which (if any) could have led to the following clustering (to 3 decimal places of precision)

	c1	c2	c3
p1	0	1	1
p2	0	1	1
p3	1	0	1
p4	1	0	1

Answer

- Agglomerative

  C3 = {C1, C2}

Question

- Of the four algorithms we covered (K-Means, agglomerative, DBSCAN, GMMs), which (if any) could have led to the following clustering (to 3 decimal places of precision)

	c1	c2	c3
p1	0	1	1
p2	0	1	1
p3	1	0	1
p4	0	0	0

Answer

- None
  - What are the properties of this clustering?

Question

- We learned about the agglomerative clustering algorithm in detail.
- In what context have we learned that divisive is also useful?

Answer

- A method for K-Means initialization!

Question

- Agree/Disagree + WHY!?
  - It is not an effective strategy to evaluate using a function over the data/clusters that is different than that optimized by your clustering algorithm

Answer

- Disagree!
  - K-Means: SSE vs Silhouette (amongst others)

Question

- Agree/Disagree
  - K-Means is guaranteed to converge and, when it does, the clustering is globally optimum.

Answer

- Disagree!
  - Converge, yes, but to a ***local*** optimum

Question

- Describe the meaning and context of the following expression in context of K-Means, where **x** refers to the dataset and **r** the one-hot membership variables

$$\frac{\sum_n r_{nk}\mathbf{x_n}}{\sum_n r_{nk}}$$

Answer

- The M-Step!!

Question

- Agree/Disagree
  - Assuming K, # of dimensions, and N are large enough, it is reasonable to assume that a small number of random Forgy restarts will result in good K-Means clusterings

Answer

- Disagree!
  - The more likely there is a bad distribution w.r.t. centroids/clusters

Question

- Increasing K will generally lead to…
  a) Monotonically higher SSE
  b) Better clusterings

Answer

😡

Question

- Agree/Disagree
  - Identifying isolated branches is a reasonable approach to finding good cluster-number cutoffs

Answer

- Disagree!
  - – Think about relative dendogrammatic heights

Question

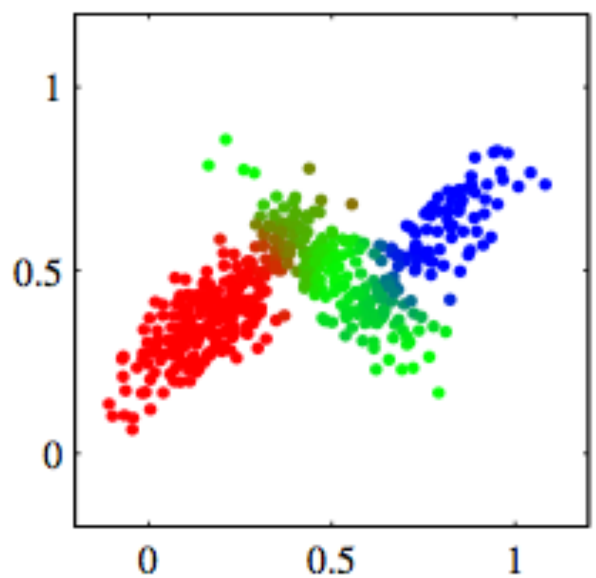
- Agree/Disagree
  - A point, q, is density-reachable from another point, p, if q is in the eps-neighborhood of p, $N_\varepsilon(p)$

Answer

- Disagree
  - What condition is missing about p?

Question

- What algorithm(s) could have produced the following clustering, where color (RGB) indicates cluster membership

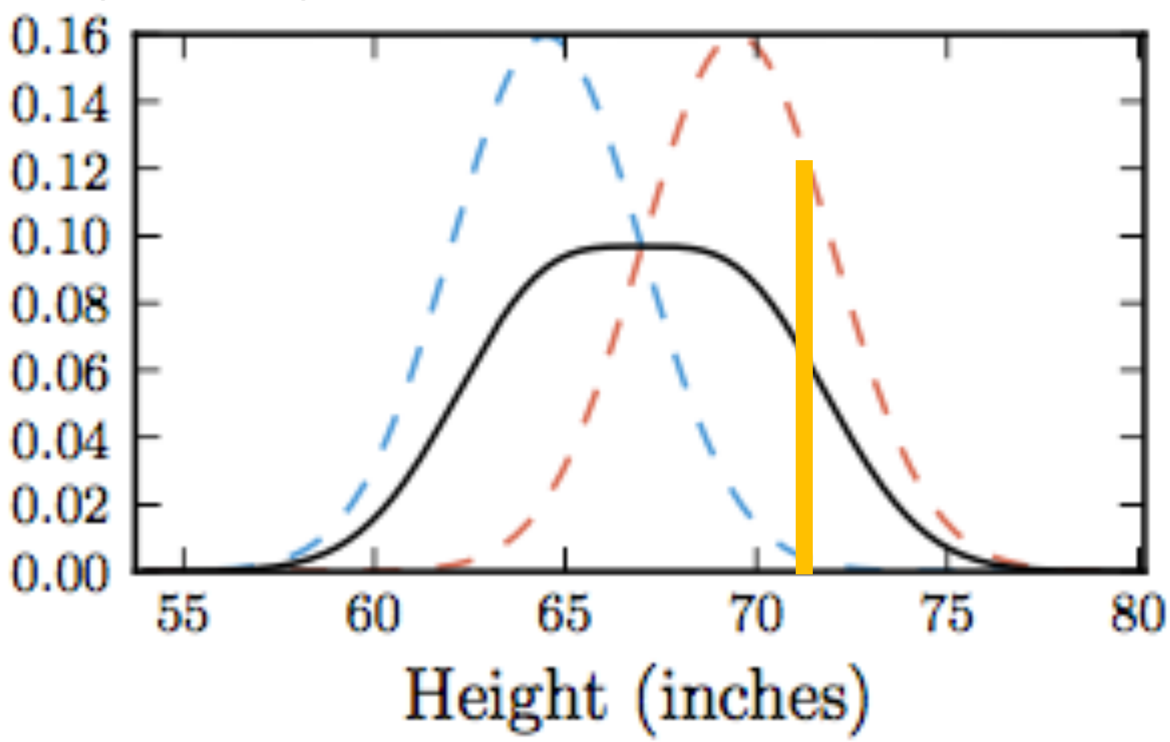


Answer

- GMMs!

Question

- Provide approximate values of responsibility for each distribution in the following diagram (close to 0/1, middle)



Answer

- Red: close to 1
- Blue: close to 0

Question

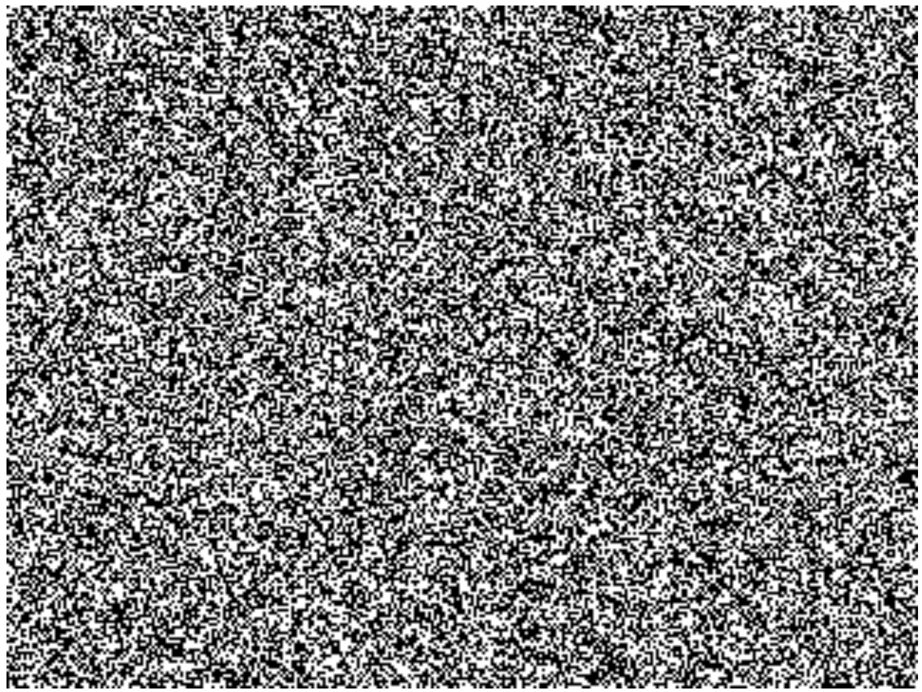
- Agree/Disagree
  - When running EM for GMMs, you stop iterating once all the responsibilities have stopped changing

Answer

- Disagree
  - Think ε

Good Clustering?

Proximity Matrix



Question

- Agree/Disagree
  - High accuracy implies high precision and/or recall

Answer

- Disagree

	Same Class	Diff Class
Same Cluster	0	25
Diff Cluster	0	175