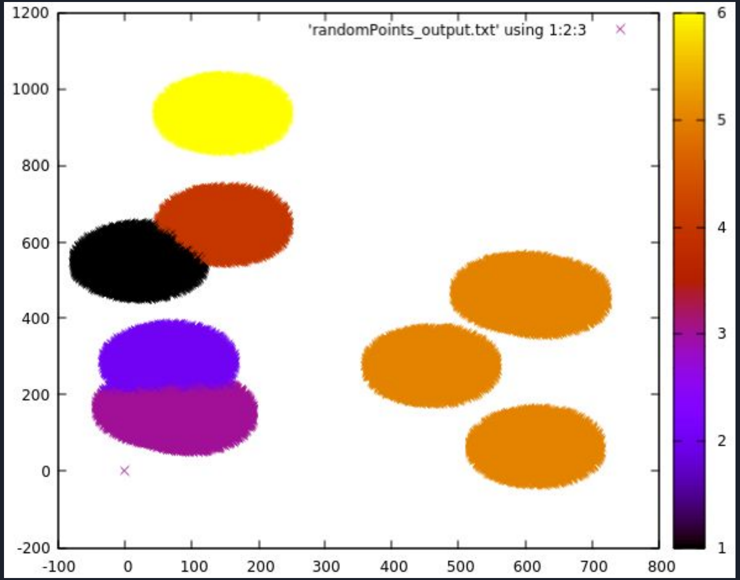
Cluster Validation

Knowing you got it right…

The Purpose of Clustering:

- “…clustering analyzes data objects without consulting class labels.” Data Mining: Concepts and Techniques.
- Associates unlabeled (unclassified) data into groups based on some distance measure.
- Distance, proximity, close-ness and similarity all measure roughly the same thing and there are many ways to measure distance.
- There are *many* decisions to make when clustering data.

Cluster Validity:



6-Means clustering of 100,000 points generated around 10 random centers.

- How can we tell that a clustering is “good”?
- Judgement is easy in low dimensional data but much harder when visualization is more difficult.
- Validation methods can either be external (which rely on external knowledge) or internal (which only relies on the distance measures within the data)

External Validation:

- Uses domain expertise to confirm good clustering.
- So to validate a clustering of students we might ask the Office of Admissions, or a clustering of animals we might ask a biologist.
- So the clustering from the previous page could be judged “bad” because I generated it with 10 centers and clustered it with 6.
- Most common but least “interesting”. Does not generate new information about data and can only validate assumptions that were already made.

Internal Validation:

- “Using this approach of cluster validity the goal is to evaluate the clustering result of an algorithm using only quantities and features inherited from the data set.” [1]
- Does not rely on any domain experts and can offer new insight into the data.
- Generally produce a numerical index which corresponds to the “goodness” of clustering.
- Methods to explore: Dunn Index, Hubert Gamma statistic, Davies-Bouldin

[1] “Cluster Validity Methods: Part 1”, Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannia, SIGMOD 2002

Dunn Index:

- First proposed by JC Dunn in the Journal of Cybernetics in 1974.
- Indicates the ratio between the minimum distance between any two clusters and the the maximum of the mean distance between points in any cluster.
- High value indicates well separated clusters.
- Because it relies on a “max value in the denominator” it is very sensitive to one bad cluster.
- □(Ci, Cj) is the inter-cluster distance measure.
    - Center to center for this presentation.
- Δk is the max distance between 2 points is a cluster.
    - Can be other measures.

$$DI_m = \frac{\min\limits_{1 \leqslant i < j \leqslant m} \delta(C_i, C_j)}{\max\limits_{1 \leqslant k \leqslant m} \Delta_k}$$

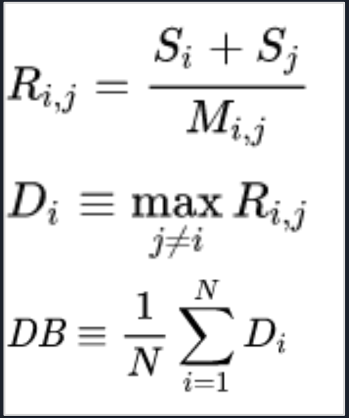
Hubert Gamma Statistic:

- Defined by S. Theodoridis and K. Koutroubas in *Pattern Recognition* in 1999.
- Indicates the sum of the products between point distances and cluster center distances.
- Value indicates compact clusters, but bigger isn’t necessarily better.
- P(i, j) is the distance between i and j.
- Q(i, j) is the distance between the centers of i’s and j’s clusters.
- M = (n*(n-1))/2

$$\Gamma = \left(1/M\right)\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} P(i, j)\cdot Q(i, j)$$

Davies-Bouldin Index:

- Defined by David Davies and Donald Bouldin in 1979 in *IEEE Transactions on Pattern Analysis and Machine Intelligence.*
- Represents a ratio between distances between points to centroids and centroids to centroids.
- Low values indicate strongly dissimilar clusters.
- S_i is the sum of all distances from a point to its own centroid i.
- $M_{i,j}$ is the distance between cluster i and j.
- N is the number of clusters.

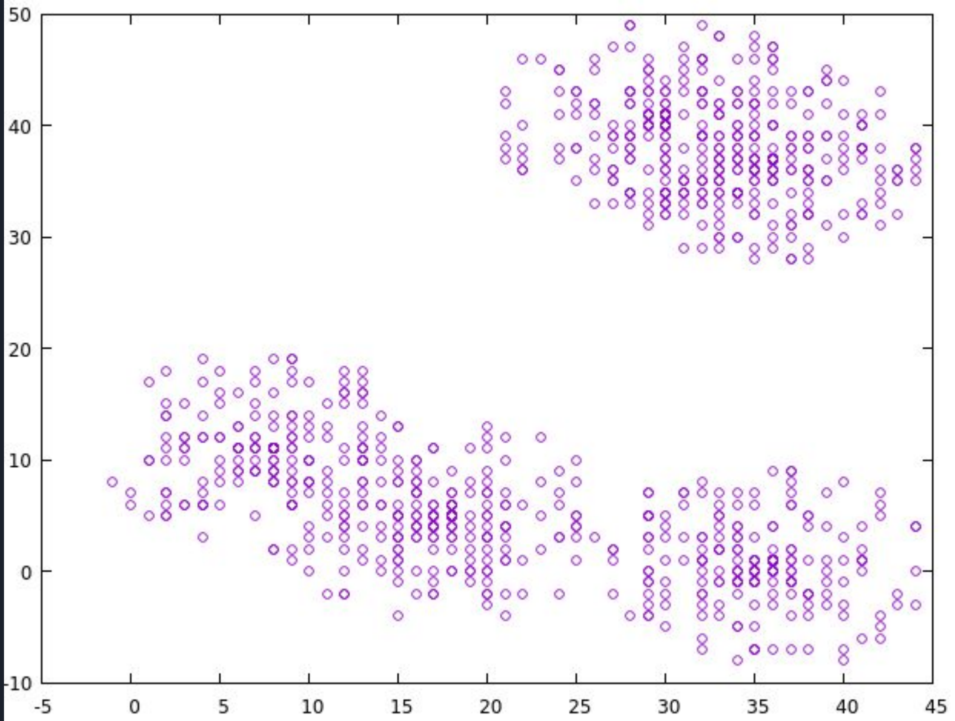


$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$$

$$D_i \equiv \max_{j \neq i} R_{i,j}$$

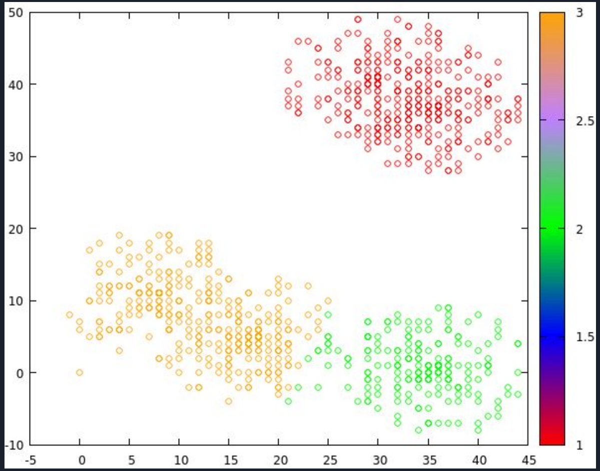
$$DB \equiv \frac{1}{N} \sum_{i=1}^{N} D_i$$

Test Data:

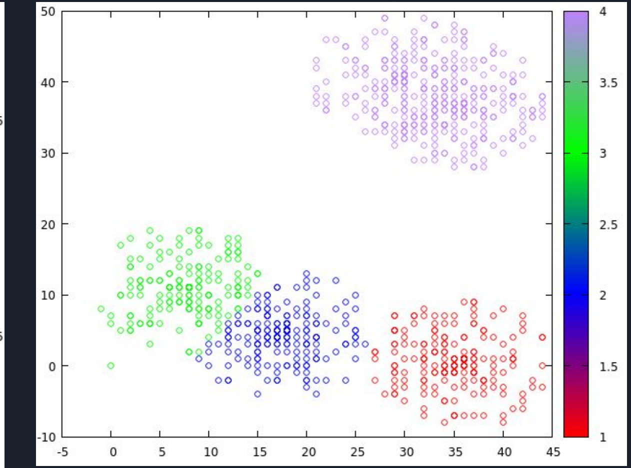


- 1000 points distributed equally at random around 5 randomly chosen centers.
- 2-dimensional
- Plotted using gnuplot

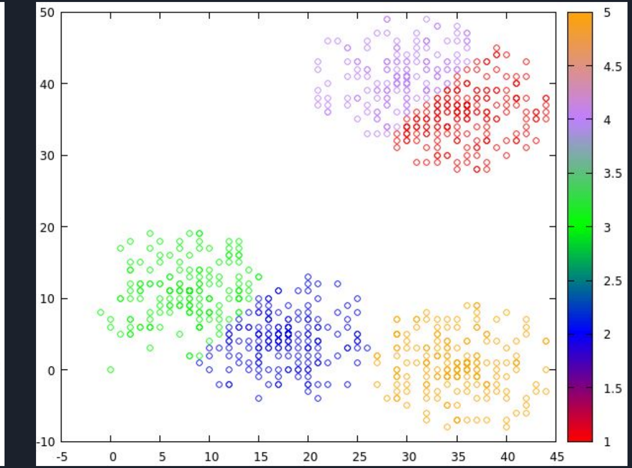
Test Clusters:



3-means

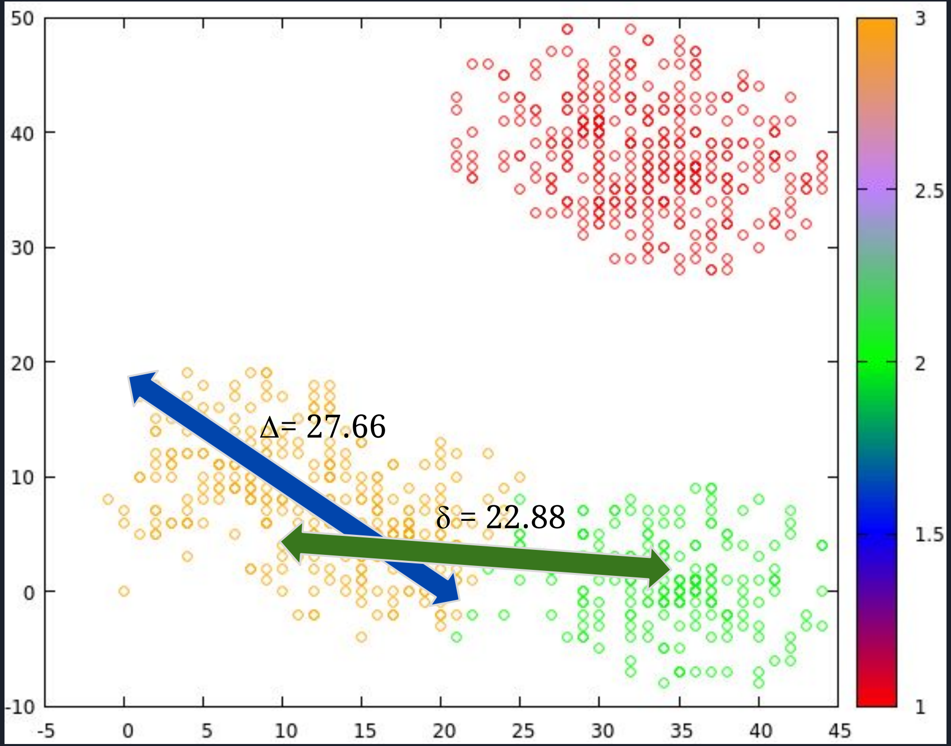


4-means



5-means

Dunn Index Results:



Largest Cluster:

27.66

Min Cluster Distance;

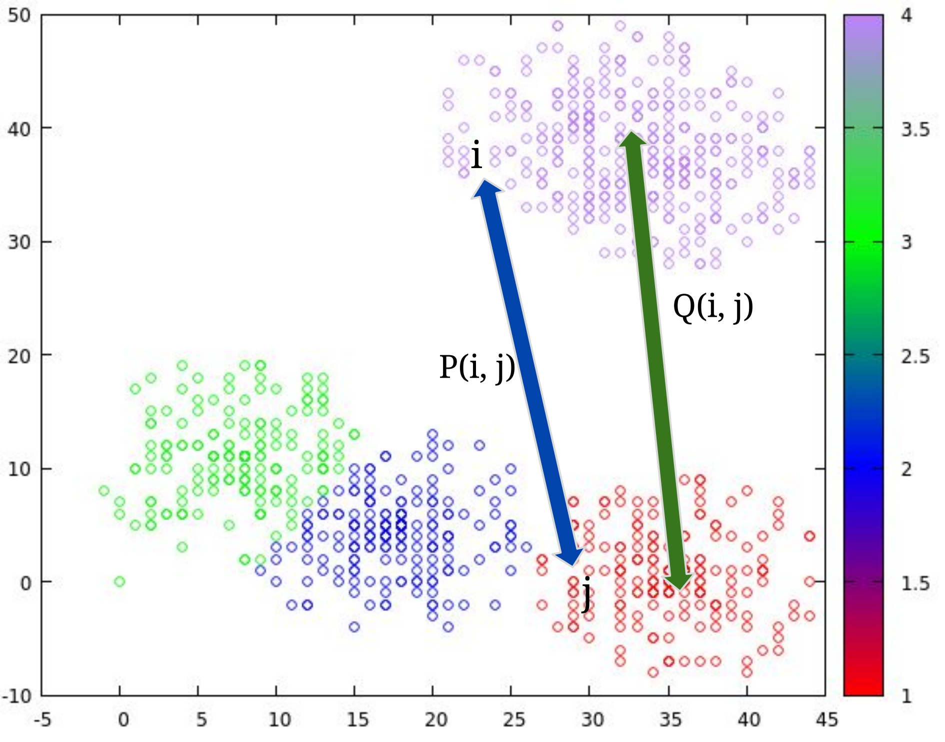
22.88

3 Clusters: 0.827

4 Cluster: 0.478

5 Clusters: 0.462

Hubert Gamma Results:



P -> point to point distance

Q -> center to center distance

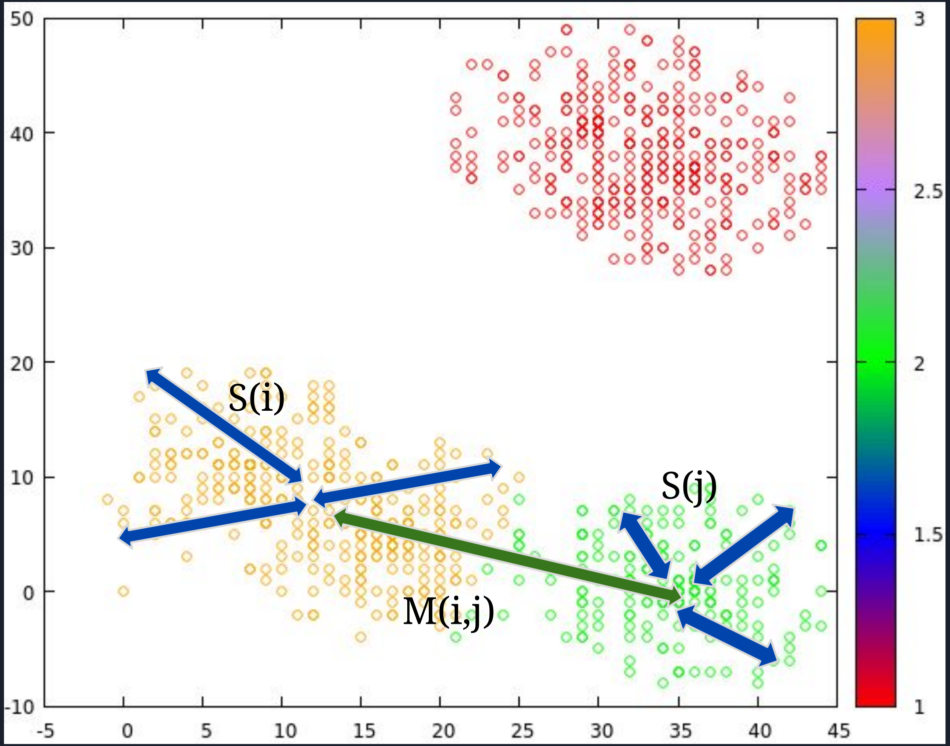
Bigger or smaller values do not necessarily indicate anything, rather we look for a “knee”.

3 clusters: 604.7

4 clusters: 620.3

5 clusters: 708.7

Davies-Bouldin Results:



S -> the distance from points to center

M -> the distance from center to center

R = (Si + Sj) / M = the similarity of two clusters

DB = the average of each cluster max R. So a smaller DB value is “better”.

3 Clusters: 0.288

4 Clusters: 0.359

5 Clusters: 0.432

Collected Results:

	3 Clusters	**4 Clusters**	**5 Clusters**
Dunn	0.827	0.478	0.462
Hubert Gamma	604.7	620.4	708.7
Davies-Bouldin	0.287	0.359	0.432
	Best separation, Most dissimilar	Best compactness	

Based on these measures, 3 clusters is our best choice.

Questions or Comments?

DOGBERT CONSULTS
CUSTOMER DATA IS AN ASSET THAT YOU CAN SELL.
IT'S TOTALLY ETHICAL BECAUSE OUR CUSTOMERS WOULD DO THE SAME THING TO US IF THEY COULD.
SOUNDS FAIR.
IN PHASE ONE, WE'LL DEHUMANIZE THE ENEMY BY CALLING THEM "DATA."
Dilbert.com DilbertCartoonist@gmail.com
10-13-10 ©2010 Scott Adams, Inc./Dist. by UFS, Inc.

Code and test data available at: https://github.com/gentry-atkinson/cs7312_assignments