Unsupervised Measures: Cohesion and Separation

- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
  - Cohesion is measured by the within cluster sum of squares (SSE)

$$SSE = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

  - Separation is measured by the between cluster sum of squares

$$SSB = \sum_{i} |C_i|(m - m_i)^2$$

Where $|C_i|$ is the size of cluster i

هرداده را از میانگین کلاستری که توشه کم میکنیم و به توان ۲ میرسونیم و این کار را به ازای تمام داده های توی کلاسترهای مختلف انجام میدیم

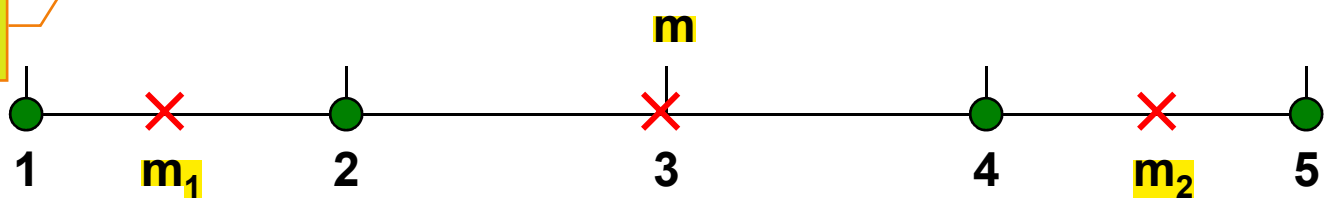
میانگین هرخوشه را از میانگین کل داده ها کم میکنیم و به توان ۲ میرسونیم و در تعداد داده های اون کلاستر ضرب میکنیم

Unsupervised Measures: Cohesion and Separation

- Example: SSE
  - SSB + SSE = constant

separation square error between clusters

SSE پراکندگی داده ها را حول میانگین داده های توی یک کلاستر اندازه میگیره که هرچی کمتر باشه نشون میده پراکندگی کمتره و داده ها به میانگین نزدیک تر هستند.



K=1 cluster:

$$SSE = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$SSB = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

فاصله ی بین خوشه ها هم بیشتر شده پس در کل از نتیجه ی این دوتا معیار میفهمیم k=2 بهتراست

K=2 clusters:

$$SSE = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$SSB = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

جمعشون ثابت میشه

وقتی k=2 باشه
sse=1 میشه یعنی پراکندگی داده های هرکلاستر نسبت به میانگین داده های توی کلاسترکمتر شده (از ۱۰ شده ۱)
پس افزایش تعداد کلاستر از یک به دو باعث بهترشدن معیار ارزیابی ما شده

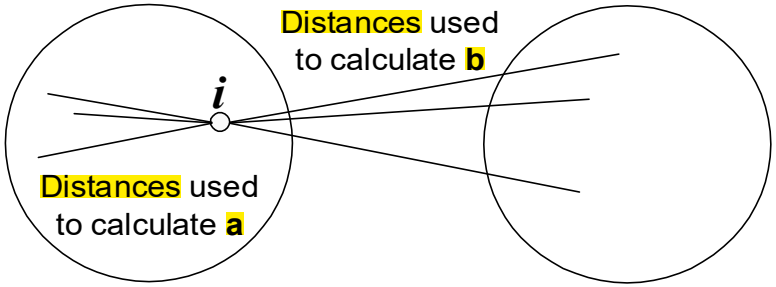
Unsupervised Measures: Silhouette Coefficient

- Silhouette coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, ***i***
  - Calculate ***a*** = average distance of ***i*** to the points in its cluster
  - Calculate ***b*** = min (average distance of ***i*** to points in another cluster)
  - The silhouette coefficient for a point is then given by

فاصله ی یک نقطه ازمیانگین داده های توی خوشه های دیگه چقدر تفاوت داره با فاصله نقطه با میانگین داده ها توی خوشه خودش

$$s = (b - a) / \max(a,b)$$

برای نرمال کردن این معیار



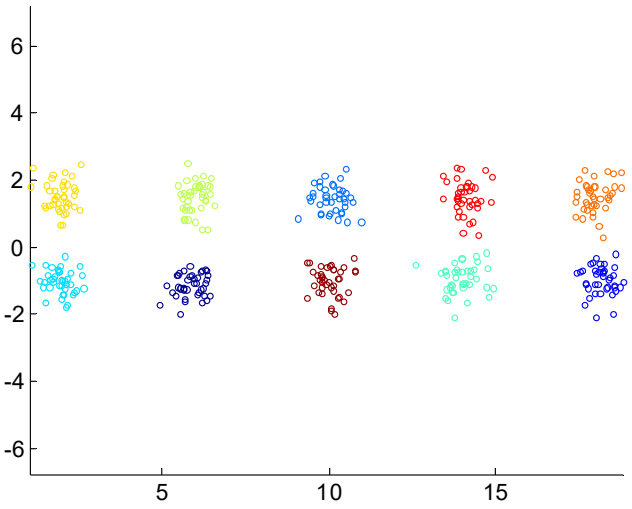
  - Value can vary between -1 and 1
  - Typically ranges between 0 and 1.
  - The closer to 1 the better.

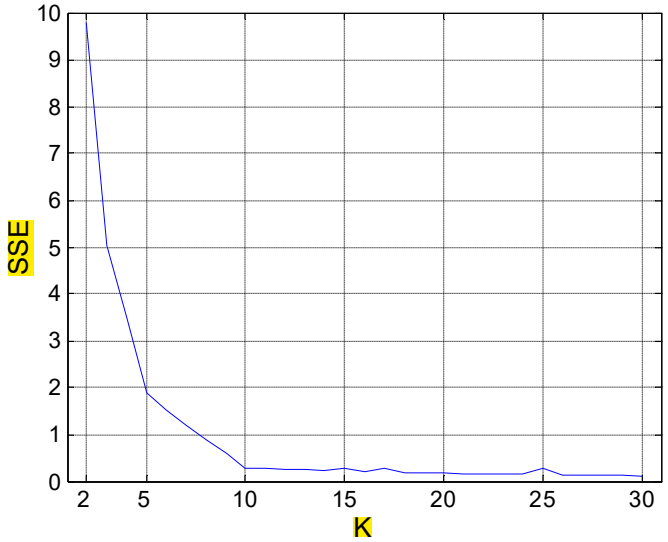
- Can calculate the average silhouette coefficient for a cluster or a clustering

Determining the Correct Number of Clusters

- SSE is good for comparing two clusterings or two clusters
- SSE can also be used to estimate the number of clusters

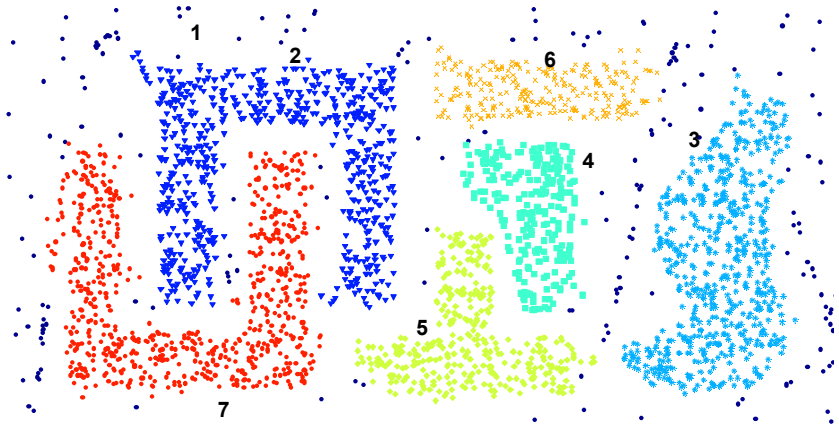
برای اینکه ممکن است روش خوشه بندی ما درمینیمم های محلی گیر کرده باشد به دلیل مقادیر اولیه مرکزها، این روش را چندبار تکرار میکنیم و نتیجه ی چندبار تکرار را میگذاریم تصمیمی که باید بگیریم

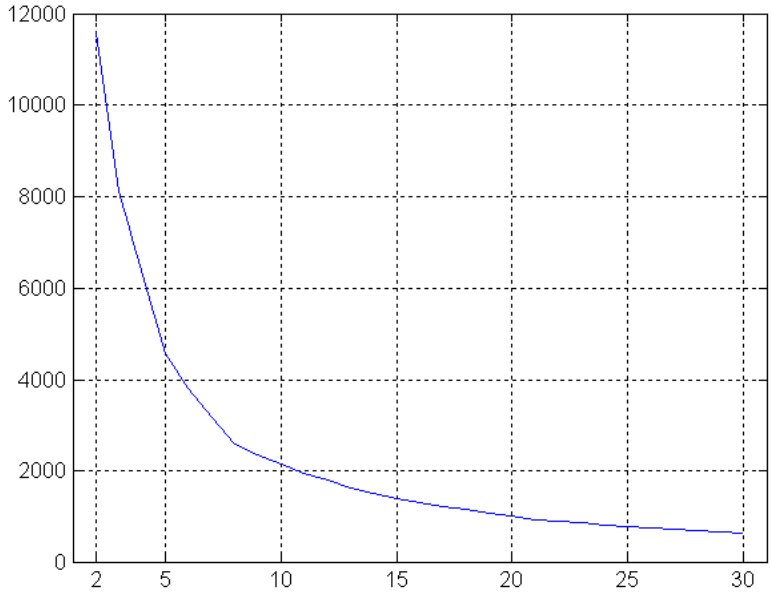




Determining the Correct Number of Clusters

- SSE curve for a more complicated data set

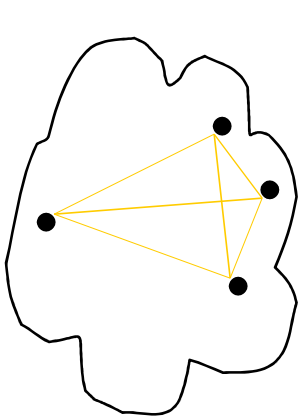




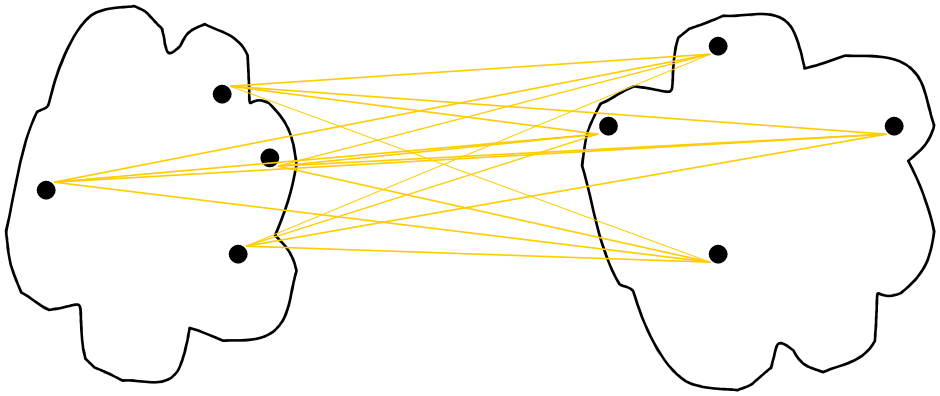
SSE of clusters found using K-means

Unsupervised Measures: Cohesion and Separation

- A proximity graph-based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion



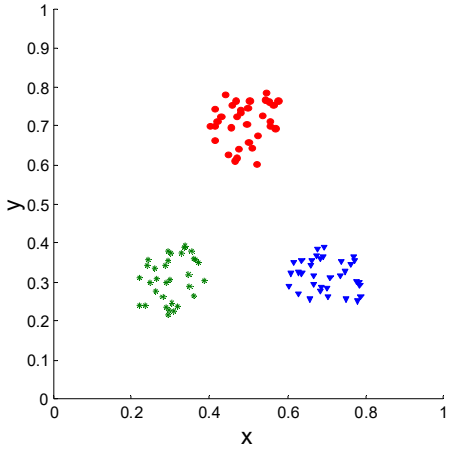
separation

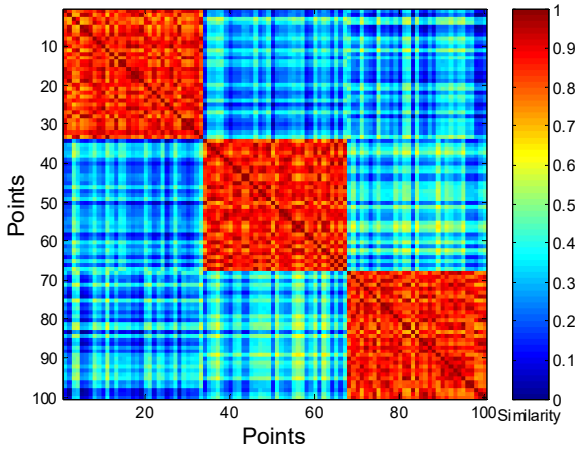
Measuring Cluster Validity Via Correlation

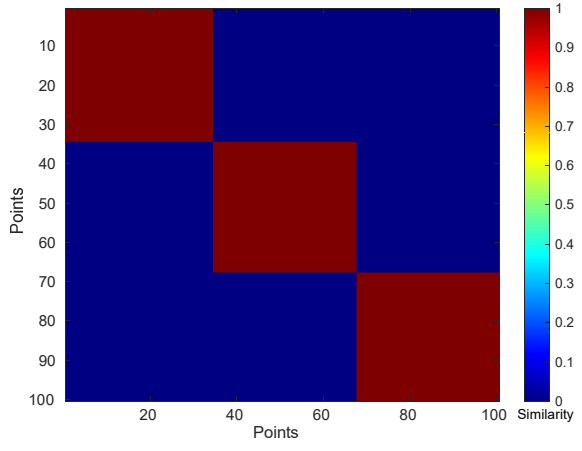
- Two matrices
  - Proximity Matrix
  - Ideal Similarity Matrix
    - One row and one column for each data point
    - An entry is 1 if the associated pair of points belong to the same cluster
    - An entry is 0 if the associated pair of points belongs to different clusters
- Compute the correlation between the two matrices
  - Since the matrices are symmetric, only the correlation between n(n-1) / 2 entries needs to be calculated.
- High magnitude of correlation indicates that points that belong to the same cluster are close to each other.
  - Correlation may be positive or negative depending on whether the similarity matrix is a similarity or dissimilarity matrix
- Not a good measure for some density or contiguity based clusters.

Measuring Cluster Validity Via Correlation

- Correlation of ideal similarity and proximity matrices for the K-means clusterings of the following well-clustered data set.



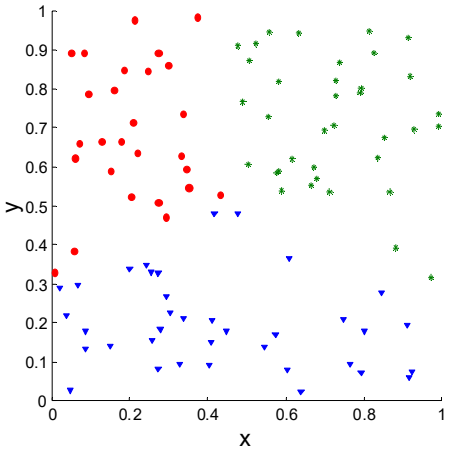


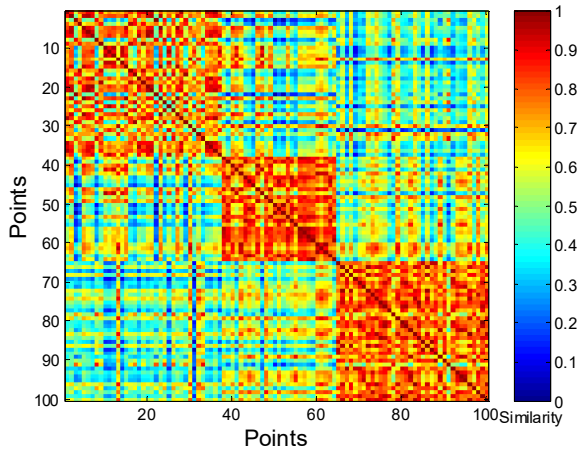


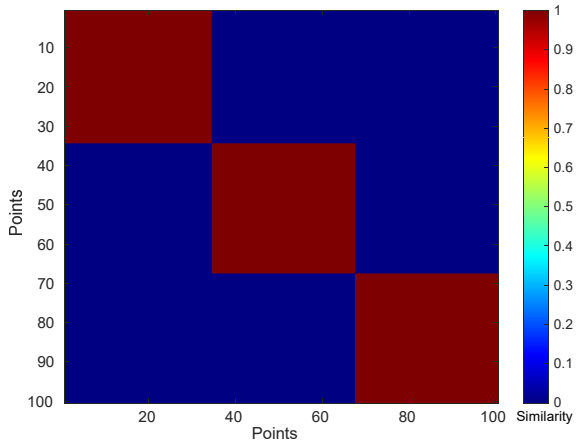
Corr = 0.9235

Measuring Cluster Validity Via Correlation

- Correlation of ideal similarity and proximity matrices for the K-means clusterings of the following random data set.





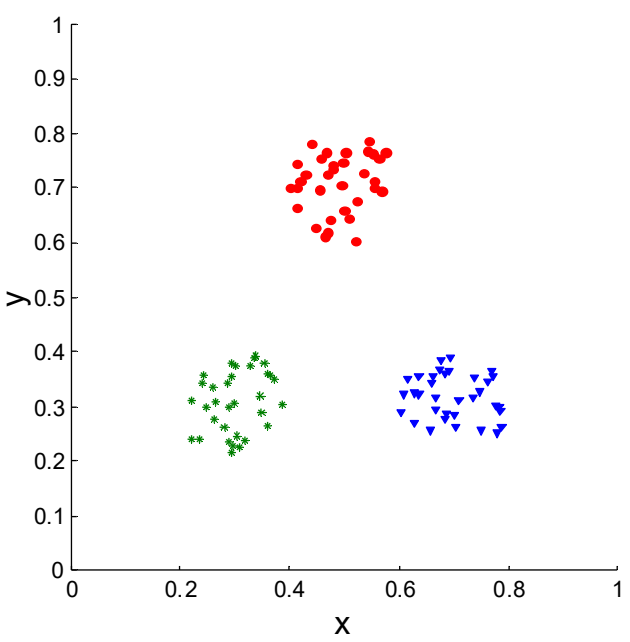


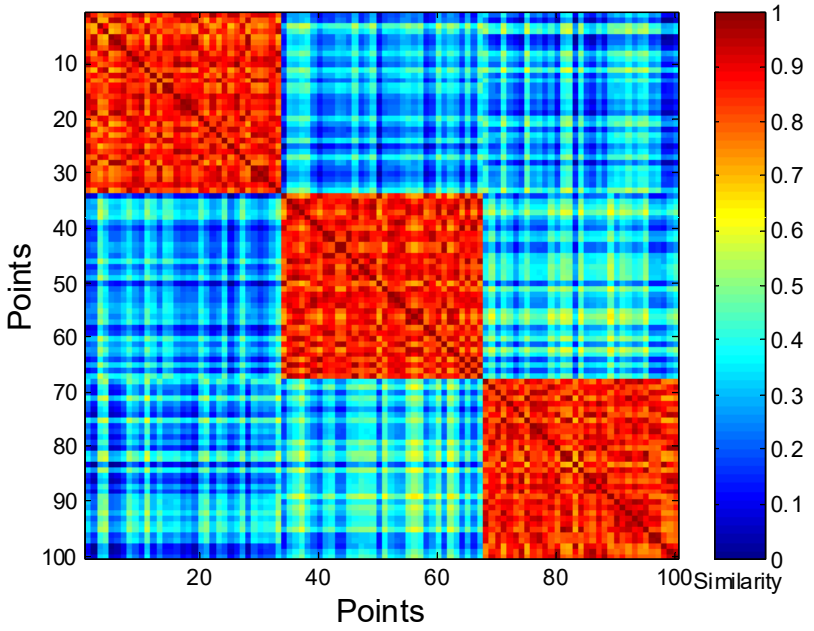
K-means

Corr = 0.5810

Judging a Clustering Visually by its Similarity Matrix

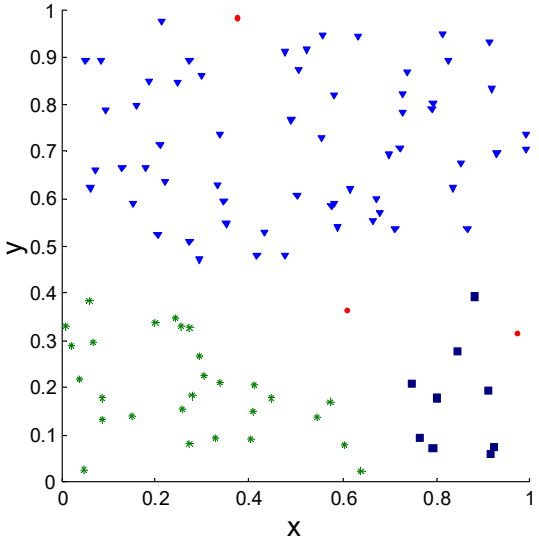
- Order the similarity matrix with respect to cluster labels and inspect visually.

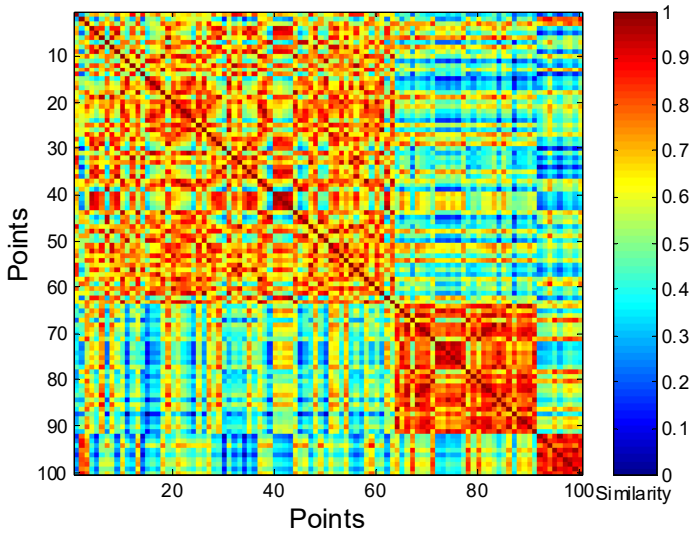




Judging a Clustering Visually by its Similarity Matrix

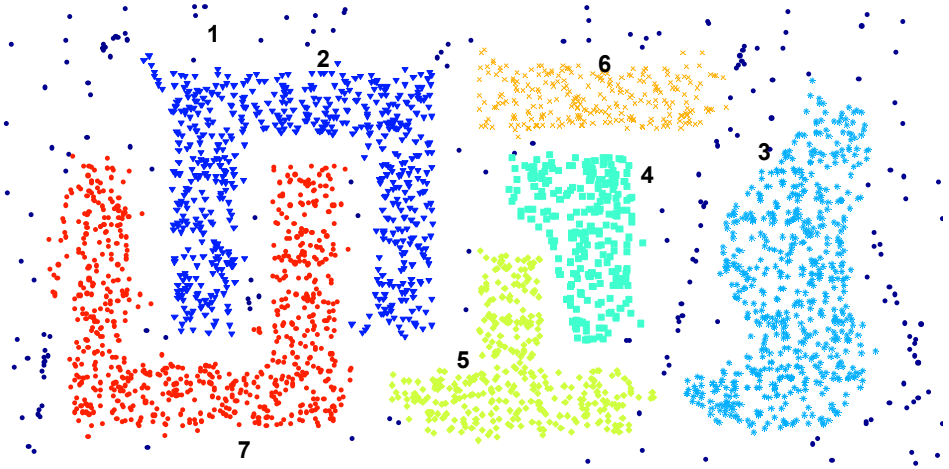
- Clusters in random data are not so crisp

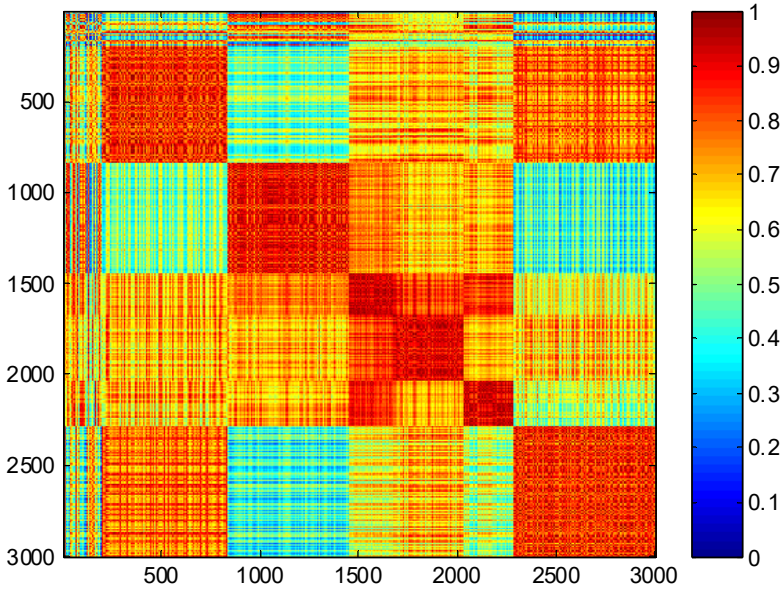




DBSCAN

Judging a Clustering Visually by its Similarity Matrix





DBSCAN

Supervised Measures of Cluster Validity: Entropy and Purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

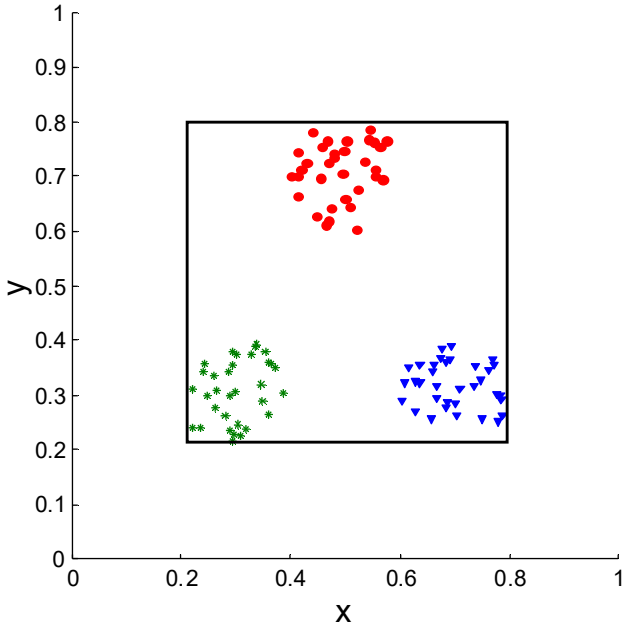
purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Assessing the Significance of Cluster Validity Measures

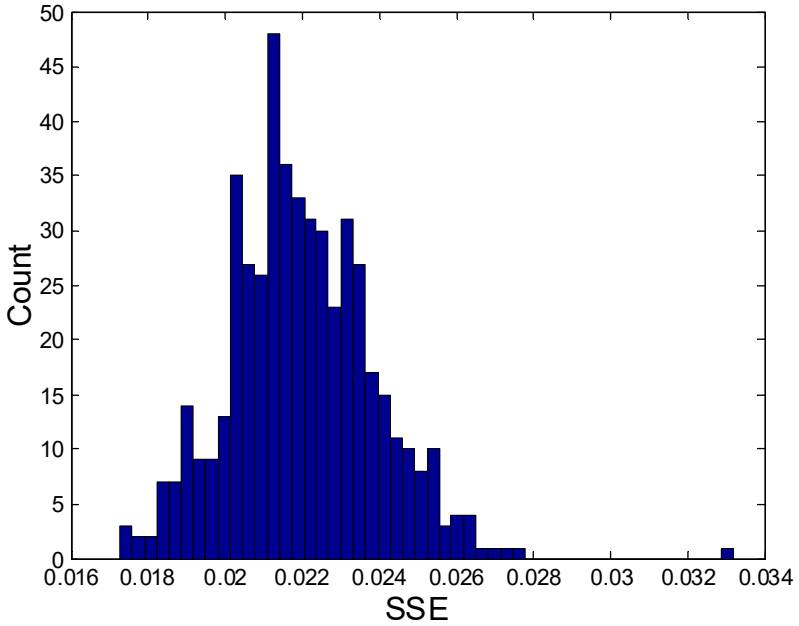
- Need a framework to interpret any measure.
  - For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - The more “atypical” a clustering result is, the more likely it represents valid structure in the data
  - Compare the value of an index obtained from the given data with those resulting from random data.
    - If the value of the index is unlikely, then the cluster results are valid

Statistical Framework for SSE

- Example
  - Compare SSE of three cohesive clusters against three clusters in random data



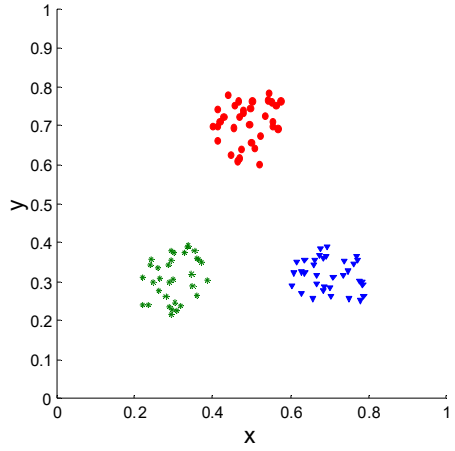
SSE = 0.005

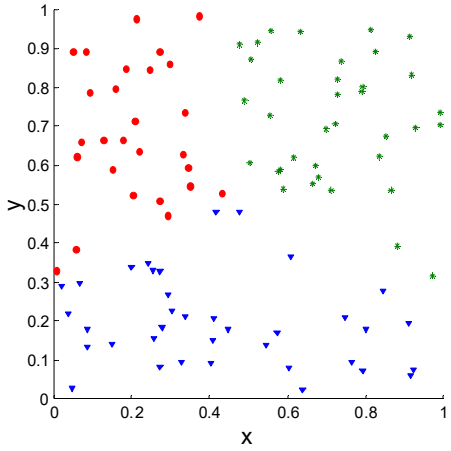


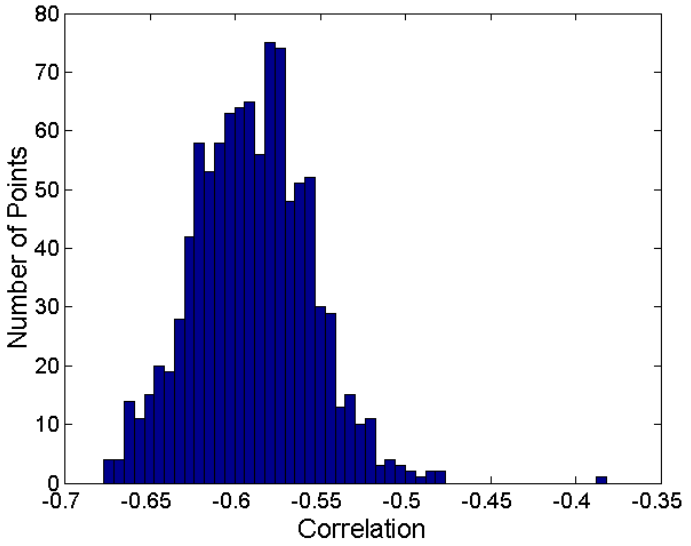
Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values

Statistical Framework for Correlation

- Correlation of ideal similarity and proximity matrices for the K-means clusterings of the following two data sets.







Corr = -0.9235

Corr = -0.5810

Histogram of correlation for 500 random data sets of size 100 with *x* and *y* values of points between 0.2 and 0.8.

Correlation is negative because it is calculated between a distance matrix and the ideal similarity matrix. Higher magnitude is better.

Final Comment on Cluster Validity

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

***Algorithms for Clustering Data*, Jain and Dubes**

- H. Xiong and Z. Li. *Clustering Validation Measures*. In C. C. Aggarwal and C. K. Reddy, editors, Data Clustering: Algorithms and Applications, pages 571–605. Chapman & Hall/CRC, 2013.