Clustering V

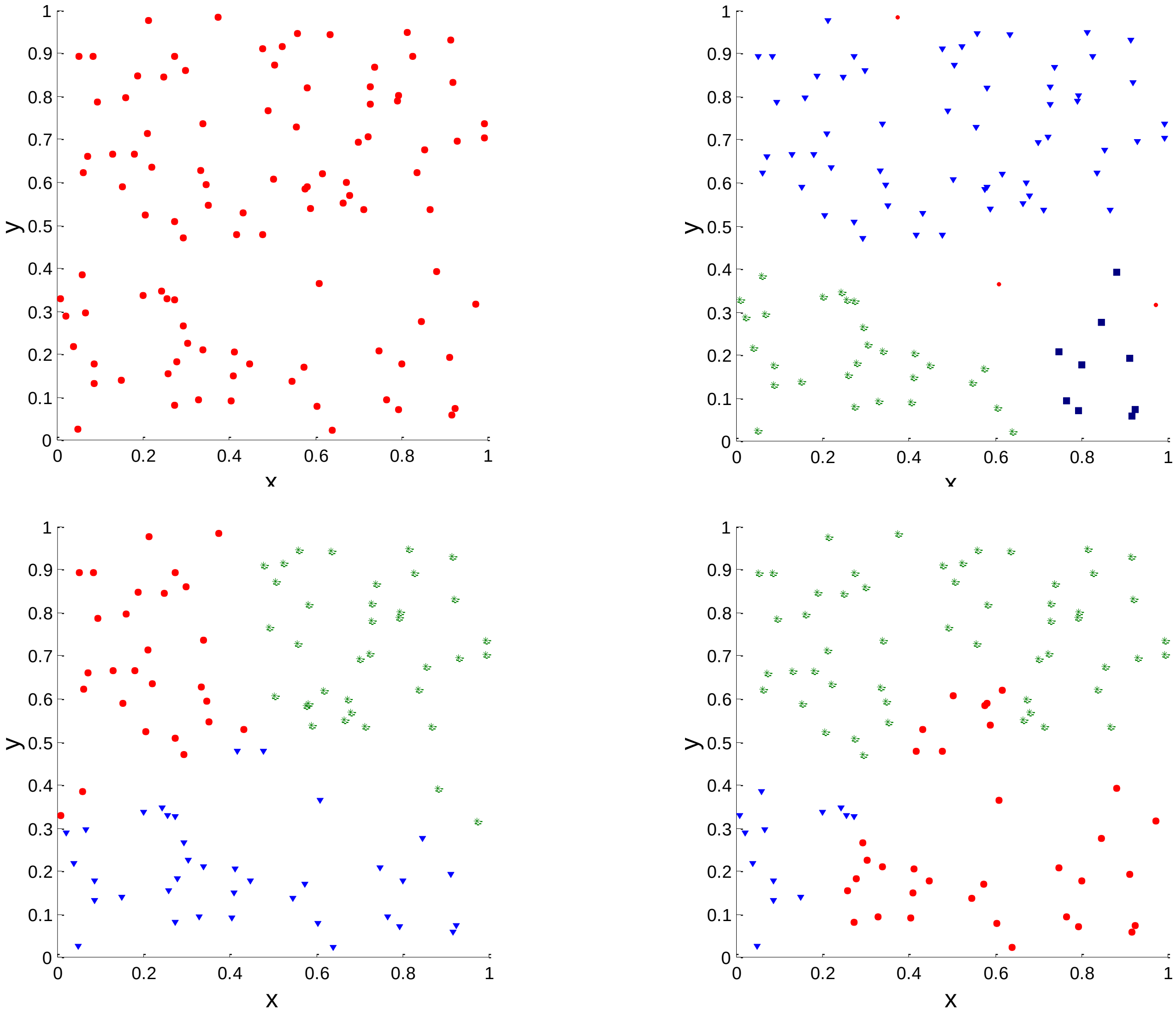
Outline

- Validating clustering results
- Randomization tests

Cluster Validity

- All clustering algorithms provided with a set of points output a clustering
- How to evaluate the “goodness” of the resulting clusters?
- Tricky because “clusters are in the eye of the beholder”!
- Then why do we want to evaluate them?
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters
  - To decide whether there is noise in the data

Clusters found in Random Data



Use the objective function **F**

- Dataset **X**, Objective function **F**
- Algorithms: $\mathbf{A_1, A_2, \ldots A_k}$
- **Question:** Which algorithm is the best for this objective function?

- $\mathbf{R_1 = A_1(X), R_2 = A_2(X), \ldots, R_k = A_k(X)}$
- Compare $\mathbf{F(R_1), F(R_2), \ldots, F(R_k)}$

Evaluating clusters

- Function **H** computes the cohesiveness of a cluster (e.g., smaller values larger cohesiveness)
- Examples of cohesiveness?
- Goodness of a cluster **c** is **H(c)**
- **c** is better than **c'** if **H(c) < H(c')**

Evaluating **clusterings** using cluster cohesiveness?

- For a clustering **C** consisting of **k** clusters $c_1,\ldots,c_k$
- $H(C) = \Phi_i H(c_i)$
- What is **Φ** ?

Cluster separation?

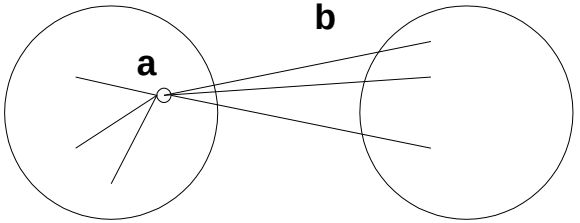
- Function **S** that measures the separation between two clusters c_i, c_j
- Ideas for **$S(c_i,c_j)$?**
- How can we measure the goodness of a clustering **$C = \{c_1,\ldots,c_k\}$** using the separation function **S**?

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*
  - ***a*** = average distance of ***i*** to the points in the same cluster
  - ***b*** = min (average distance of ***i*** to points in another cluster)
  - silhouette coefficient of **i**:

    s = 1 – a/b if **a < b**

  - Typically between 0 and 1.
  - The closer to 1 the better.



- Can calculate the Average Silhouette width for a cluster or a clustering

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

Assessing the significance of clustering (and other data mining) results

- Dataset **X** and algorithm **A**
- Beautiful result **A(D)**
- But: what does it mean?
- How to determine whether the result is really interesting or just due to chance?

Examples

- Pattern discovery: frequent itemsets or association rules
- From data **X** we can find a collection of nice patterns
- Significance of individual patterns is sometimes straightforward to test
- What about the whole collection of patterns? Is it surprising to see such a collection?

Examples

- In clustering or mixture modeling: we always get a result
- How to test if the whole idea of components/clusters in the data is good?
- Do they really exist clusters in the data?

Classical methods – Hypothesis testing

- **Example:** Two datasets of real numbers **X** and **Y** (**|X|=|Y|=n**)
- **Question:** Are the means of **X** and **Y** (resp. **E(X)**, **E(Y)**) are significantly different

- Test statistic: **t = (E(X) – E(Y))/s**, (**s**: an estimate of the standard deviation)

- The test statistic follows (under certain assumptions) the **t** distribution with **2n-2** degrees of freedom

Classical methods – Hypothesis testing

- The result can be something like: **"the difference in the means is significant at the level of 0.01"**
- That is, if we take two samples of size **n**, such a difference would occur by chance only in about 1 out of 100 trials
- **Problems:**
  - What if we are testing many hypotheses (multiple hypotheses testing)
  - What if there is no closed form available?

Classical methods: testing independence

X	Y
1	1
1	1
1	1
1	1
1	1
1	1
1	0
1	0
0	1
0	1
0	0

- Are columns **X** and **Y** independent?
- **Independence**: **Pr(X,Y) = Pr(X)*Pr(Y)**
  - Pr(X=1) = 8/11, Pr(X=0)=3/11, Pr(Y=1) = 8/11, Pr(Y=0) = 3/11
  - **Actual joint probabilities**: Pr(X=1,Y=1) = 6/11, Pr(X=1,Y=0)=2/11, Pr(X=0,Y=1) = 2/11, Pr(X=0.Y=0)=1/11
  - **Expected joint probabilities:** Pr(X=1,Y=1) = 64/121, Pr(X=1,Y=0)=24/121, Pr(X=0,Y=1) = 24/121, Pr(X=0,Y=0)=9/121

Testing independence using χ^2

X	Y
1	1
1	1
1	1
1	1
1	1
1	1
1	0
1	0
0	1
0	1
0	0

- Are columns **X** and **Y** independent?

	Y=1	Y=0	∑row
X=1	6	2	8
Y=0	2	1	3
∑column	8	3	11

$$\chi^2 = \sum_{\{x,y\}\in\{0,1\}^2} \frac{E[X=x,Y=y]-O(X=x,Y=y)^2}{E[X=x,Y=y]}$$

- **So what?**

Classical methods – Hypothesis testing

- The result can be something like: **“the independence between X and Y is significant at the level of 0.01”**

- That is, if we take two columns X and Y with the observed P(X=1) and P(Y=1) and **n** rows, such degree of independence would occur by chance only in about 1 out of 100 trials

Problems with classical methods

- What if we are testing many hypotheses (multiple hypotheses testing)

- What if there is no closed form available?

Randomization methods

- **Goal:** assessing the significance of results
  - Could the result have occurred by chance?

- **Methodology:** create datasets that somehow reflect the characteristics of the true data

Randomization methods

- Create randomized versions from the data **X**
- **$X_1, X_2, \ldots, X_k$**
- Run algorithm **A** on these, producing results **$A(X_1), A(X_2), \ldots, A(X_k)$**
- Check if the result **A(X)** on the real data is somehow different from these
- ***Empirical p-value:*** the fraction of cases for which the result on real data is (say) larger than **A(X)**
- If the empirical **p**-value is small, then there is something interesting in the data

Randomization for testing independence

X	Y
1	1
1	1
1	1
1	1
1	1
1	1
1	0
1	0
0	1
0	1
0	0

- **P_x = Pr(X=1) and P_y = Pr(Y=1)**
- Generate random instances of columns **(X_i,Y_i)** with parameters **P_x** and **P_y** [independence assumption]
- **p-value:** Compute the in how many random instances, the **χ^2** statistic is greater/smaller than its value in the input data

Randomization methods for other tasks

- Instantiation of randomization for clustering?

- Instantiation of randomization for frequent-itemset mining

Columnwise randomization: no global view of the data

X	Y
1	1
1	1
1	1
1	1
1	1
1	1
1	0
1	0
0	1
0	1
0	0
0	0

X	Y
1	1
1	1
1	1
1	1
1	1
1	1
1	0
1	0
0	1
0	1
0	0
0	0

Columnwise randomization: no global view of the data

X	Y	...
1	1	0 0 1 0 0 1 1
1	1	1 1 0 0 1 0 0
1	1	0 0 0 1 0 1 1
1	1	0 1 1 0 1 0 1
1	1	0 1 0 0 0 0 1
1	1	1 0 1 0 0 1 0
1	0	0 0 0 1 1 0 0
1	0	0 1 1 0 0 0 1
0	1	0 0 1 1 0 0 0
0	1	1 0 0 1 0 0 1
0	0	0 0 1 0 1 0 0
0	0	0 1 1 0 1 0 0

X	Y	...
1	1	1 1 1 1 1 1 1
1	1	1 1 1 1 1 1 1
1	1	0 1 1 1 1 1 1
1	1	1 1 1 1 1 1 1
1	1	1 1 1 1 0 1 1
1	1	1 1 1 1 1 1 1
1	0	
1	0	
0	1	
0	1	
0	0	
0	0	

X and **Y** are not more surprisingly correlated given that they both have **1**s in dense rows and **0**s in sparse rows

Questions

- What is a good way of randomizing the data?
- Can the sample $\mathbf{X_1, X_2, \ldots, X_k}$ be computed efficiently?
- Can the values $\mathbf{A(X_1), A(X_2), \ldots, A(X_k)}$ be computed efficiently?

What is a good way of randomizing the data?

- How are datasets X_i generated?

- What is the underlying ***"null model"/ "null hypothesis"***

Swap randomization

- **0—1** data: **n** rows, **m** columns, presence/absence

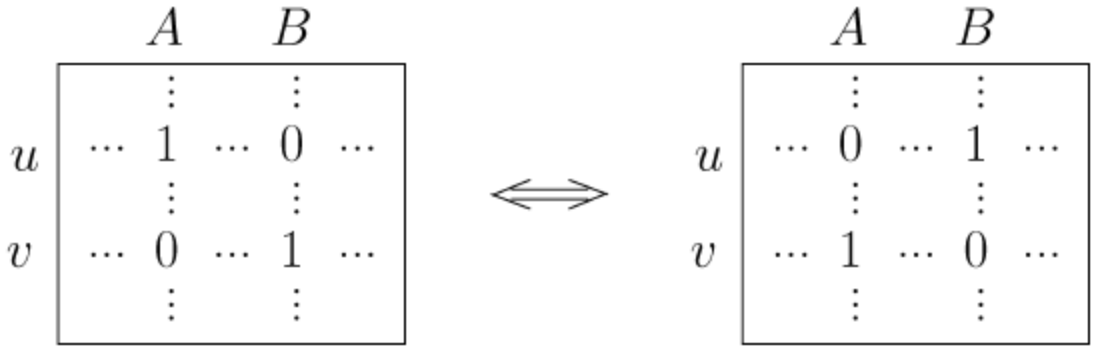
- Randomize the dataset by generating random datasets with the ***same row and column margins*** as the original data

- Reference: A. Gionis, H. Mannila, T. Mielikainen and P. Tsaparas: Assessing data-mining results via swap randomization (TKDD 2006)

Basic idea

- Maintains the degree structure of the data

- Such datasets can be generated by ***swaps***



Fixed margins

- ***Null hypothesis:*** the row and the column margins of the data are fixed

- If the marginal information is known, then what else can you say about the data?

- What other structure is there in the data?

Example

X	Y	...
1	1	0 0 1 0 0 1 1
1	1	1 1 0 0 1 0 0
1	1	0 0 0 1 0 1 1
1	1	0 1 1 0 1 0 1
1	1	0 1 0 0 0 0 1
1	1	1 0 1 0 0 1 0
1	0	0 0 0 1 1 0 0
1	0	0 1 1 0 0 0 1
0	1	0 0 1 1 0 0 0
0	1	1 0 0 1 0 0 1
0	0	0 0 1 0 1 0 0

Significant co-occurrence of **X** and **Y**

X	Y	...
1	1	1 1 1 1 1 1 1
1	1	1 1 1 1 1 1 1
1	1	0 1 1 1 1 1 1
1	1	1 1 1 1 1 1 1
1	1	1 1 1 1 0 1 1
1	1	1 1 1 1 1 1 1
1	0	0 0 0 1 1 0 0
1	0	0 1 1 0 0 0 1
0	1	0 0 1 1 0 0 0
0	1	1 0 0 1 0 0 1
0	0	0 0 1 0 1 0 0

No significant co-occurrence of **X** and **Y**

Swap randomization and clustering

Dataset	k	E	mean	std	Z	p
S1	10	1777.3	3669.9	11.1	170.43	0.01
S2	10	4075.4	4084.4	11.6	0.77	0.22
COURSES	10	17541.6	24405.1	30.2	227.09	0.01
PALEO	10	1040.7	1401.7	4.8	74.74	0.01
RETAIL	10	23920.9	24086.0	135.2	1.22	0.10

Error in clustering
Of the real data

Mean error in clustering
of the swapped data