Machine Learning for NLP: Unsupervised learning techniques

Saturnino Luz

Dept. of Computer Science, Trinity College Dublin, Ireland

ESSLLI'07 ○ Dublin ○ Ireland

Supervised vs. unsupervised learning

- So far we have seen supervised learning (of classification):
  - learning based on a training set where labelling of instances represents the target (categorisation) function
  - classifier implements an approximation of the target funtion
  - outcome: a classification decision
- Unsupervised learning:
  - learning based on unannotated instances;
  - outcome: a grouping of objects (instances and groups of instances)

Notes

Applications

- Exploratory data analysis (data mining): clustering can reveal patterns of association in the data
- Information visualisation: natural ways of displaying association patterns
  - dendrograms, self-organising maps etc
- Information retrieval: keyword [Sparck Jones and Jackson, 1970] and document [van Rijsbergen, 1979] clustering.
- Improving language models
- Corpus analysis (homogeneity)
- Object and character recognition
- *Dimensionality reduction by term extraction* in text categorisation

Data representation

- As before, vector-based representation is a popular choice. E.g.:

		lecture	we	examined	clustering	groups	...	
lecture	$= \langle$	2,	2,	1,	2,	0	, ...	$\rangle$
we	$= \langle$	2,	2,	1,	2,	0	, ...	$\rangle$
examined	$= \langle$	1,	1,	1,	2,	0	, ...	$\rangle$
clustering	$= \langle$	2,	2,	1,	3,	1	, ...	$\rangle$
groups	$= \langle$	0,	0,	0,	1,	1	, ...	$\rangle$
$\vdots$				$\vdots$				

Figure 1: Co-occurrence vector representation for words

Notes

4-1

Types of unsupervised learning

- Clustering algorithms are the main technique for unsupervised learning;
- A taxonomy [Jain et al., 1999]:
  - Partitional clustering:
    - k-$means$, Expectation Maximisation (EM), Graph theoretic, mode seeking
  - hierarchical:
    - single-link
    - complete-link
    - average-link
  - Agglomerative vs. divisive

Notes

5-1

Distance and dissimilarity measures

- Given instances a, b and c represented as real-valued vectors, a *distance* between a and b is a function $d(a, b)$ satisfying:

$$
\begin{aligned}
d(a,b) &\geq 0 && (1)\\
d(a,a) &= 0 && (2)\\
d(a,b) &= d(b,a) && (3)\\
d(a,b) &\leq d(a,c) + d(b,c) && (4)
\end{aligned}
$$

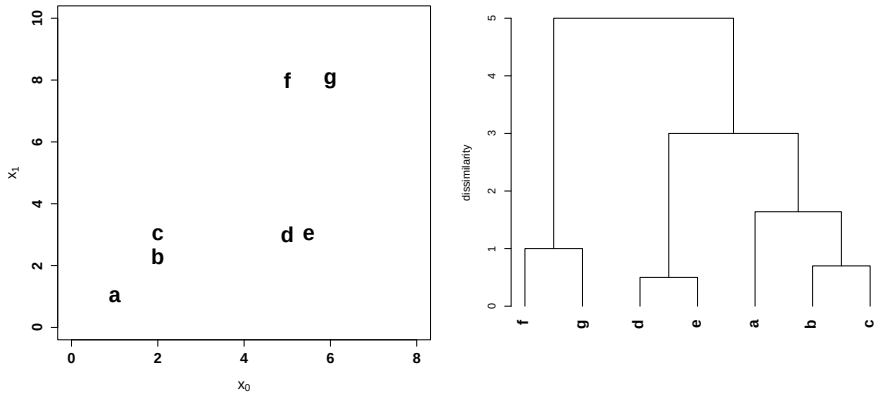
- When (4) doesn't hold, d is called a *dissimilarity*
- Euclidean distance, $d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{|\vec{x}|}(x_i - y_i)^2}$ is commonly used.

Notes

6-1

Hierarchical clustering

- Input: objects represented as vectors
- Output: a hierarchy of associations represented as a "dendrogram"



(If you know R, see hclusters.R in ronaldo.cs.tcd.ie/esslli07/practicals/)

Notes

7-1

A simple agglomerative clustering algorithm

Algorithm 1: Simple agglomerative hierarchical clustering

```
1  hclust(D: set of instances): tree
2    var: C,  /* set of clusters */
3         M  /* matrix containing distances between */
4            /* pairs of clusters                    */
5    for each d ∈ D} do
6       make d a leaf node in C
7    done
8    for each pair a,b ∈ C do
9       M_{a,b} ← d(a,b)
10   done
11   while (not all instances in one cluster) do
12     Find the most similar pair of clusters in M
13     Merge these two clusters into one cluster.
14     Update M to reflect the merge operation.
15   done
16   return C
```


Notes

8-1

Similarity

- Results vary depending on how you define similarity.
- The definition determine the type of clustering algorithm:
  - In *Single-link* clustering, similarity is defined as the *minimum* distance between any two pairs of instances:

$$sim_s(c_1, c_2) = \frac{1}{1 + \min_{x_1 \in c_1, x_2 \in c_2} d(x_1, x_2)} \quad (5)$$

  - In *complete-link*, as the *maximum* distance between any two pairs of instances:

$$sim_c(c_1, c_2) = \frac{1}{1 + \max_{x_1 \in c_1, x_2 \in c_2} d(x_1, x_2)} \quad (6)$$

  - and in *average-link*, as the mean distance:

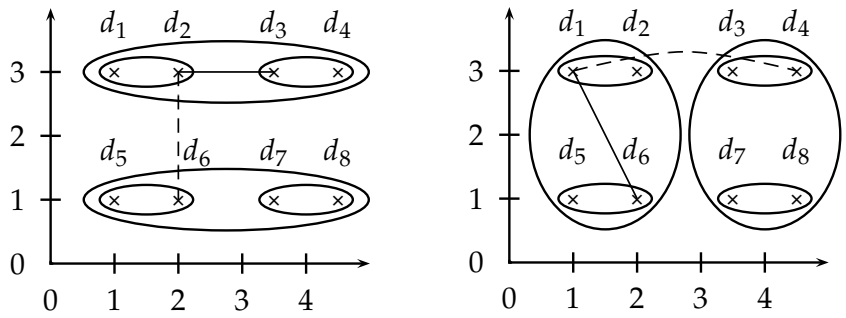
$$sim_a(c_1, c_2) = \frac{1}{1 + \frac{1}{|c_1||c_2|} \sum_{x_1 \in c_1} \sum_{x_2 \in c_2} d(x_1, x_2)} \quad (7)$$

Notes

9-1

How do the different definitions affect clustering?

- Single-link tend to produce "straggly" or elongated clusters whereas complete-link tend to produce more compact groups [Manning and Schütze, 1999]:



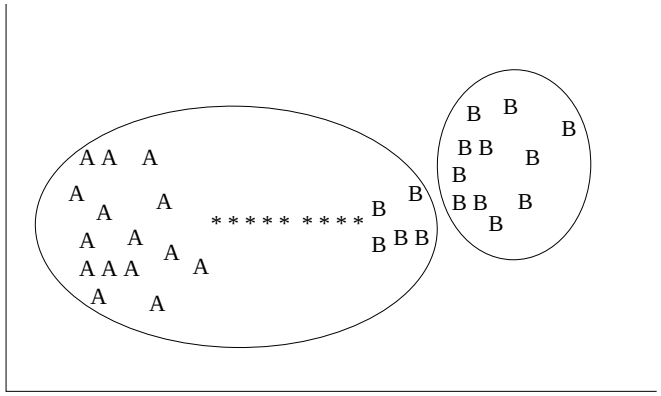
Single link Complete-link

Notes

10-1

Why are elongated clusters sometimes a bad thing?

- noise data in the vicinity of clusters might lead to incorrect merging:

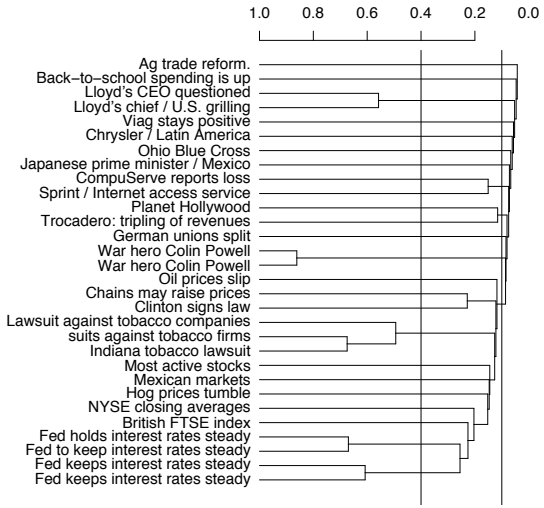


Notes

11-1

Examples: clustering of RCV1 documents

- Dendrogram for single-link clustering of 30 RCV1 documents (Manning, Raghavan & Schültze, in press):

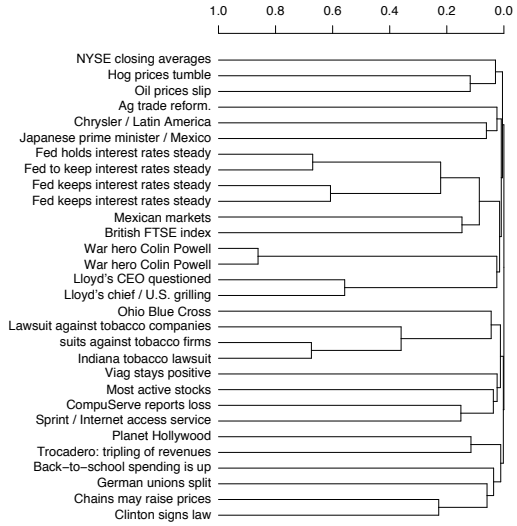


Notes

12-1

Examples: clustering of RCV1 documents

- Dendrogram for complete-link clustering of 30 RCV1 documents:



Notes

13-1

k-means clustering

Algorithm 2: K-means clustering

```
1  k-means(X = {d_1, ..., d_n} ⊆ R^m, k): 2^R
2      C : 2^R   /* μ a set of clusters */
3      d : R^m × R^m → R /* distance function */}
4      μ : 2^R → R /* μ computes the mean of a cluster */}
5    select C with k initial centres f_1, ..., f_k
6    while stopping criterion not true do
7      for all clusters c_j ∈ C do
8          c_j ← {d_i | ∀f_l d(d_i, f_j) ≤ d(d_i, f_l)}
9      done
10     for all means f_j do
11         f_j ← μ(c_j)
12     done
13   done
14   return C
```


Notes

14-1

k-means characteristics

- Need to select the number of clusters in advance
- Might converge to a local minimum
- But...
  - it is more efficient (lower computational complexity) than hirearchical clustering
- K-means can be seen as a specialisation of the expectation maximisation (EM) algorithm

Notes

15-1

Another Example: term extraction for TC

Sample co-occurrence matrix for a subset of REUTERS-21578:

usair	20	2	0	1	0	4	1	0	0	0	1	0	2	0	3	0	0	0	1	0	1	1	0	0	0	0	0	0	0	2	14	1	3
voting	2	10	0	2	0	1	0	0	0	0	0	0	2	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	2	0	0
buyout	0	0	8	1	0	2	0	0	0	0	0	0	1	1	0	0	0	0	0	3	0	0	0	0	0	0	0	1	0	1	0	0	0
stake	1	2	1	62	0	0	0	1	1	0	0	1	2	0	0	0	0	2	0	1	0	0	1	0	1	2	1	0	0	0	1	0	0
santa	0	0	0	0	7	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0	1	0
merger	4	1	2	0	3	48	0	1	3	0	2	0	1	2	4	1	0	0	1	1	2	0	0	0	2	0	0	1	0	5	4	3	1
ownership	1	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
rospatch	0	0	0	1	0	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rexnord	0	0	0	1	0	3	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
designs	0	0	0	0	0	0	0	0	0	5	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	1	0	0	0
pie	1	0	0	0	0	2	0	0	0	0	5	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	1
⋮	⋮																																

Notes

16-1

Extraction with k-means

K-means ($k = 5$) clustering of the words in slide 16

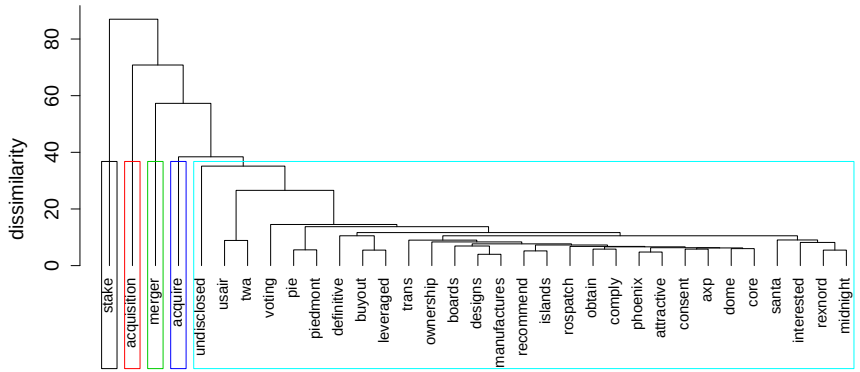
Cluster	elements
1	stake
2	usair, merger, twa
3	acquisition
4	acquire
5	voting, buyout, santa, ownership, rospatch, rexnord, designs, pie, recommend, definitive, piedmont, consent, boards, dome, obtain, leveraged, comply, phoenix, core, manufactures, midnight, islands, axp, attractive, undisclosed, interested, trans

Notes

17-1

Term extraction by hierarchical clustering

Hierarchical clustering (complete-link) of the words in slide 17



Notes

18-1

Unsupervised word-sense disambiguation

Concordances for the word "bank"

```
RIV y be? Then he ran down along the bank, toward a narrow, muddy path.
FIN  four bundles of small notes the bank cashier got it into his head
RIV ross the bridge and on the other bank you only hear the stream, the
RIV beneath the house, where a steep bank of earth is compacted between
FIN op but is really the branch of a bank. As I set foot inside, despite
FIN raffic police also belong to the bank. More foolhardy than entering
FIN require a number.  If you open a bank account, the teller identifies
RIV  circular movement, skirting the bank of the River Jordan, then turn
.                                   .
.                                   .
.                                   .
```


Notes

19-1

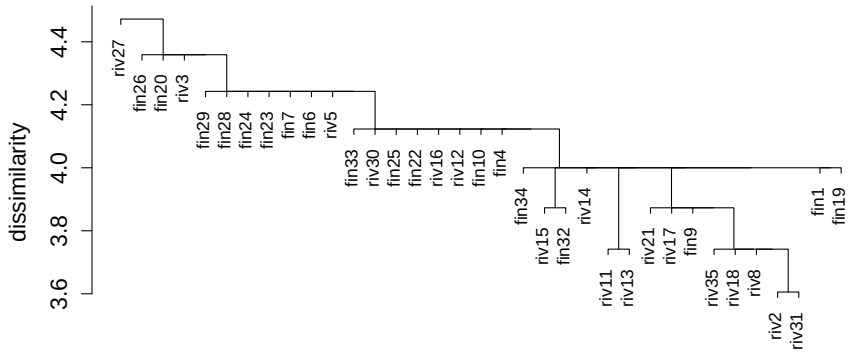
Sample run of 2-means clustering

- k-means clusters the lines into the following groups:
  1. fin1, riv3, fin4, fin6, fin7, fin9, fin10, riv15, riv16, fin19, fin20, fin22, fin23, fin24, fin25, fin26, riv27, fin28, fin29, fin32, fin33, fin34
  2. riv2, riv5, riv8, riv11, riv12, riv13, riv14, riv17, riv18, riv21, riv30, riv31, riv35

Notes

20-1

Hierarchical clustering (single-link) of senses of the word "bank"



Notes

21-1

Further topics

- Efficient clustering algorithms
- Cluster labelling
- Cluster evaluation
- Expectation Maximisation (EM) clustering and applications
- Clustering and information visualisation: SOM and ANNs

References

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999. URL `citeseer.ist.psu.edu/jain99data.html`.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

K. Sparck Jones and D.M. Jackson. The use of automatically-obtained keyword classifications for information retrieval. *Information Storage and Retrieval*, 5:175–201, 1970.

C. J. van Rijsbergen. *Information Retrieval.* Butterworths, 1979. URL `http://www.dcs.gla.ac.uk/Keith/`.

Notes