

9.1 Linear Clustering

Machine Learning 1: Foundations

Marius Kloft (TUK)

Kaiserslautern, 16–23 June 2020

Recap

So far we have required so far in this course that both is given:

① inputs $x_1, \dots, x_n \in \mathbb{R}^d$

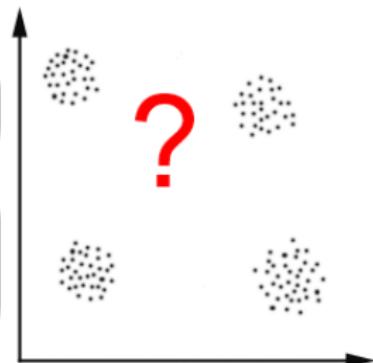
② labels y_1, \dots, y_n

- ▶ with either $y_i \in \{-1, +1\}$ (for binary classification)
- ▶ or $y_i \in \mathbb{R}$ (for regression)

This learning setting is called **supervised learning**.

But what if there are no labels given?

This setting is called
unsupervised learning.



Today and Next Week

Definition

Unsupervised learning is the area of machine learning where

- ▶ we are given inputs $x_1, \dots, x_n \in \mathbb{R}^d$
- ▶ but **no labels** y_1, \dots, y_n .

What could be interesting tasks in unsupervised learning?

1. Anomaly Detection

Detect unusual data points

- ▶ i.e., that deviate strongly from the previously seen ones

Examples:



Corrupt data



Novelties



Outliers



Rare events



Attacks on networks

2. Density Estimation and Generative Models

Learn the **law** by which nature has generated the data



And use it to **generate** new examples!



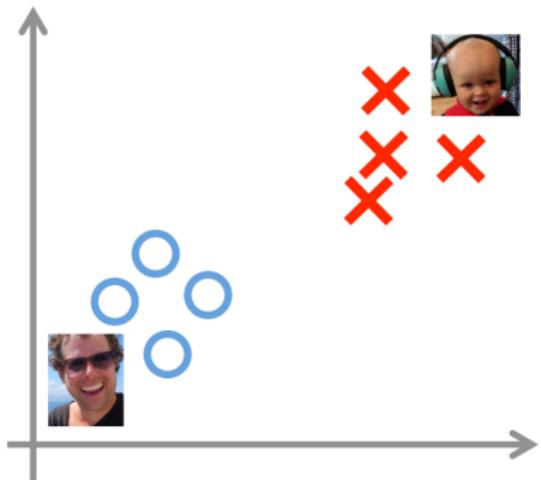
We will learn in **ML2** how this works

3. Dimensionality Reduction (**Next Week!**)

Compress data into lower-dimensional representation

Applications:

- ▶ data visualization
- ▶ data de-noising



4. Clustering

Today!

Contents of this Class

Clustering

- 1 Linear Clustering
- 2 Non-linear Clustering
- 3 Hierarchical Clustering

1 Linear Clustering

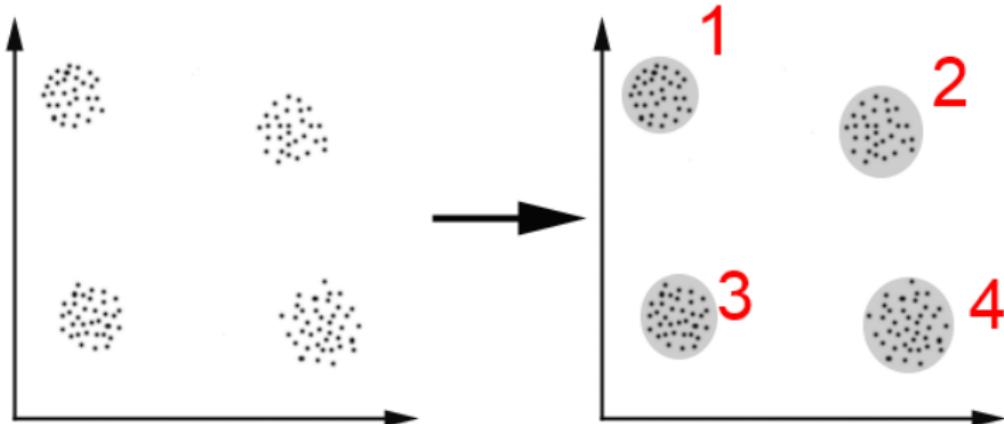
2 Non-linear Clustering

3 Hierarchical Clustering

What is Clustering?

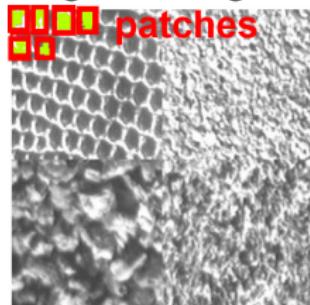
Definition

Clustering is the process of organizing objects into groups—called **clusters**—whose members are similar in some way.



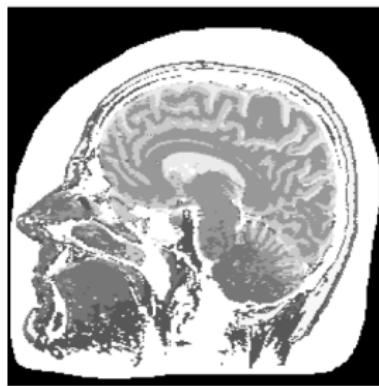
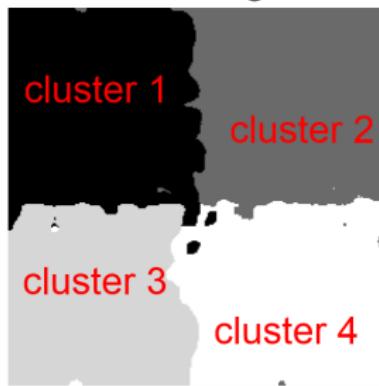
Example: Image Segmentation

Original image:



each patch is a
training example

segmented image



More examples

- ▶ **Recommender systems:** organizing products and customers into groups that are similar



- ▶ **Social networks:** cluster users into groups that have similar interests/preferences



- ▶ **WWW document classification:** organize webpages (e.g., news articles) into clusters with similar content (sports, economics, ...)

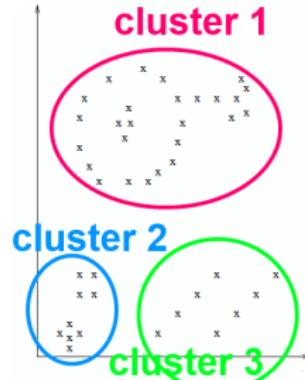
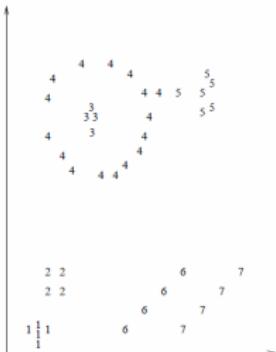
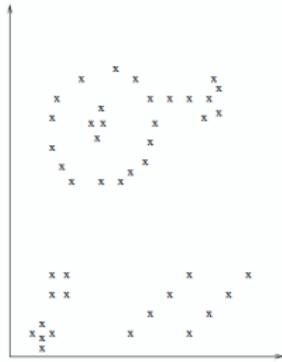


Clustering Dilemma

Big problem:

- ▶ evaluation of results!

Why? Quiz: how many clusters in this example?



7 clusters

vs.

3 clusters

Evaluation of results

- ▶ Ground truth unknown (**no labels!**)
- ⇒ Results need to be inspected manually
- ▶ Example:
 - ▶ Given collection of documents,

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

- ▶ cluster words into groups

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

cluster 1 cluster 2 cluster 3 cluster 4

| | | | |
|---------|------------|----------|------------|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

Evaluation of results

- ▶ Ground truth unknown (**no labels!**)
- ⇒ Results need to be inspected manually
- ▶ Example:
 - ▶ Given collection of documents,

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants, an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

- ▶ cluster words into groups & manual labeling

Arts Budgets Child Education

NEW
FILM
SHOW
MUSIC
MOVIE
PLAY
MUSICAL
BEST
ACTOR
FIRST
YORK
OPERA
THEATER
ACTRESS
LOVE

MILLION
TAX
PROGRAM
BUDGET
BILLION
FEDERAL
YEAR
SPENDING
NEW
STATE
PLAN
MONEY
PROGRAMS
GOVERNMENT
CONGRESS

CHILDREN
WOMEN
PEOPLE
CHILD
YEARS
FAMILIES
WORK
PARENTS
SAYS
FAMILY
WELFARE
MEN
PERCENT
CARE
LIFE

SCHOOL
STUDENTS
SCHOOLS
EDUCATION
TEACHERS
HIGH
PUBLIC
TEACHER
BENNETT
MANIGAT
NAMPHY
STATE
PRESIDENT
ELEMENTARY
HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants, an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

k-means Clustering Algorithm

Arguably the most popular clustering algorithm is the following:

```
1: function KMEANS(parameter  $k$ , inputs  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ )
2:   initialize cluster centers  $\mathbf{c}_1, \dots, \mathbf{c}_k$ 
   (e.g., randomly drawn inputs  $\mathbf{c}_j := \mathbf{x}_j$ )
3:   repeat
4:     for  $i = 1 : n$  do
5:       label the input  $\mathbf{x}_i$  as belonging to the nearest cluster,
         
$$y_i := \arg \min_{j=1, \dots, k} \|\mathbf{x}_i - \mathbf{c}_j\|^2$$

6:     end for
7:     for  $j = 1 : k$  do
8:       compute cluster center  $\mathbf{c}_j$  as the mean of all inputs of the  $j$ th cluster,
         
$$\mathbf{c}_j := \text{mean}(\{\mathbf{x}_i : y_i = j\})$$

9:     end for
10:    until convergence criterion is met
        (e.g., no change in clusters between subsequent iterations)
11:    return cluster centers  $c_1, \dots, c_k$ 
12: end function
```

Demos

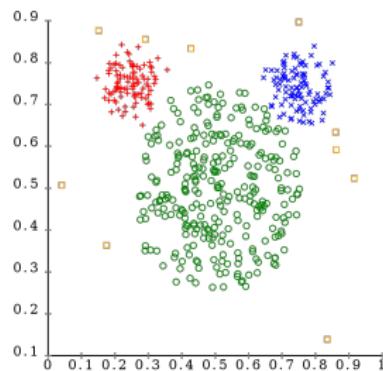
- ▶ http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html
- ▶ <http://www.cs.washington.edu/research/imagedatabase/demo/kmcluster/>
- ▶ <http://syskall.com/kmeans.js/>

Limitations:

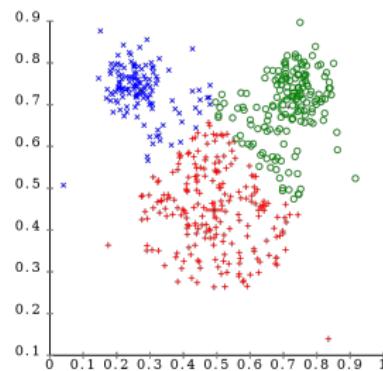
k-means Can Fail to Find the Right Clusters

Different cluster analysis results on "mouse" data set:

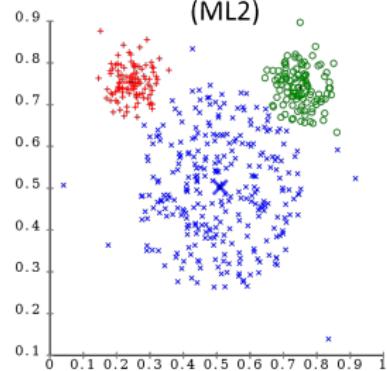
Original Data



k-Means Clustering



mixture of Gaussians
(ML2)



Will learn about an improvement of *k*-means in ML2: **mixture of Gaussians**

Limitations:

k-means Finds Linear Cluster Boundaries

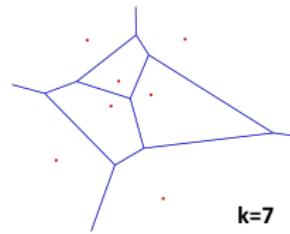
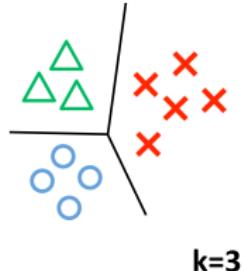
Easiest to see for $k = 2$ clusters:

- ▶ Then the cluster boundary is the one of the nearest centroid classifier



For $k \geq 3$ clusters:

- ▶ Cluster boundaries given by Voronoi diagram of cluster centers
- ▶ Thus the boundaries are (piecewise) linear



What to Do If the Ideal Decision Boundary is Non-linear?

