Unsupervised Learning

BE Chapman, PhD

August 16, 2023

What is learning?

"Learning is any process by which a system improves performance from experience." (Herbert Simon)

Why would we want a computer to learn?

- Why would we want them to learn?
- How do we get them to learn?

Two broad categories of machine learning

- Unsupervised learning
  - Algorithms operate on unlabelled examples
  - Task: Clustering
- Supervised learning
  - Algorithms are trained on labelled examples
  - Tasks: Classification & Regression
- Supervised learning is more common, but ...

Unsupervised learning

- Labels can be hard to get in healthcare
- Why?

What is clustering?

- The organization of unlabeled data into similarity groups called clusters.
- A cluster is a collection of data items which are "similar" between them, and "dissimilar" to data items in other clusters.
- Why is clustering useful?
  - Does not require labels
    - Usually costly/time consuming
    - E.g., characterising protein function experimentally
  - Good for discovering patterns
  - Discover the hidden structure of the data

What is clustering?

- Clustering is an **optimization problem**.
- What are we going to optimize?
- Dissimilarity
  - Let c_i be the ith cluster
  - Let C be your overall clustering $C = \cup c_i$

$$variability(c) = \sum_{x \in c} distance(centroid(c), e)^2$$

$$dissimilarity(C) = \sum_{c \in C} variability(c)$$

- How does variability differ from variance?

What is clustering?

- Clustering is an **optimization problem**.
- What are we going to optimize?
- Minimize dissimilarity
  - Let c_i be the ith cluster
  - Let C be your overall clustering $\cup c_i$

$$variability(c) = \sum_{x \in c} distance(centroid(c), e) dissimilarity(C) \quad = \sum_{c \in C} variability(c)$$

What is an obvious way of minimizing dissimilarity?

Clustering is optimization with constraints!

- Example constraints:
  - number of clusters
  - minimum number of elements in a cluster
  - maximum distance between elements in a cluster

Ingredients for Clustering

- Representation of the data
- Distance (inverse similarity) measure
- Quality measure
- Algorithm
  - constraints

Section 1

Representation

Vector

- We are almost always representing our data with **vectors**
- What is a vector?
  - A vector is a column of numbers
  - a row of numbers is a **transpose** of a vector ($\mathbf{x}^\mathbf{T}$

A Vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \tag{1}$$

- N is the number of entries in the vector: the **dimension** of the vector
- $N \in \{1, 2, 3, \cdots\}$

Example: 1D Vector

$$\mathbf{x_1} = [\mathbf{3}]$$
$$\mathbf{x_2} = [\mathbf{4}]$$
$$\vdots$$
$$\mathbf{x_{18}} = [\mathbf{98}]$$

Example: 1D Vector

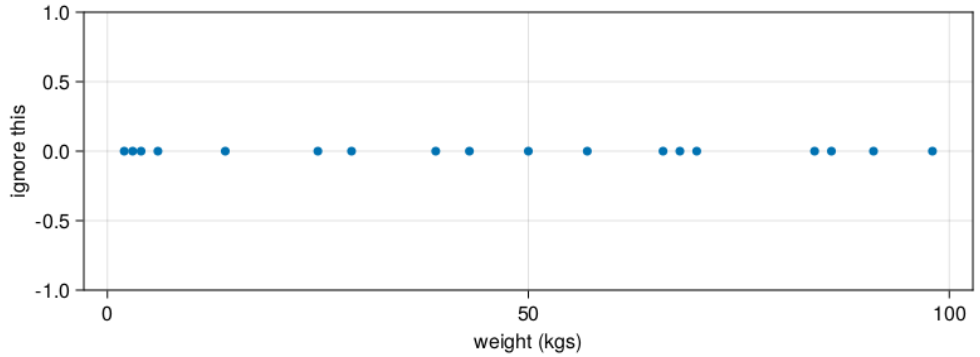


Figure 1: weight

Example: 2D Vector

$$\mathbf{x_1} = \begin{bmatrix} 3 \\ 51 \end{bmatrix}$$

$$\mathbf{x_2} = \begin{bmatrix} 4 \\ 46 \end{bmatrix}$$

$$\mathbf{x_{18}} = \begin{bmatrix} 98 \\ 203 \end{bmatrix}$$

Example: 2D Vector

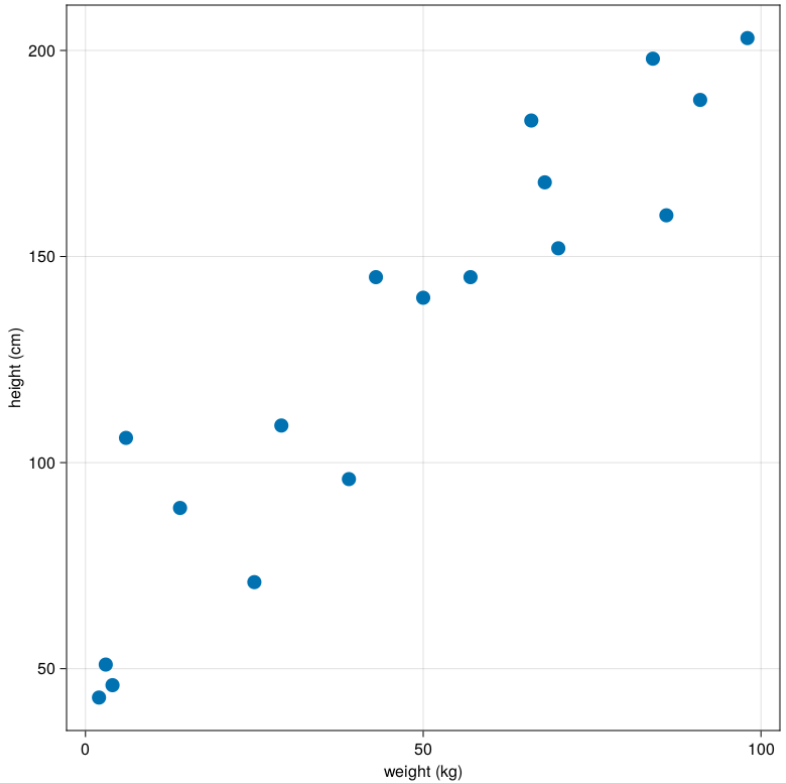


Figure 2: height and weight

Example: 3D Vector

$$\mathbf{x_1} = \begin{bmatrix} 3 \\ 51 \\ 0.0 \end{bmatrix}$$

$$\mathbf{x_2} = \begin{bmatrix} 4 \\ 46 \\ 0.1 \end{bmatrix}$$

$$\mathbf{x_{18}} = \begin{bmatrix} 98 \\ 203 \\ 55 \end{bmatrix}$$

Example: 3D Vector

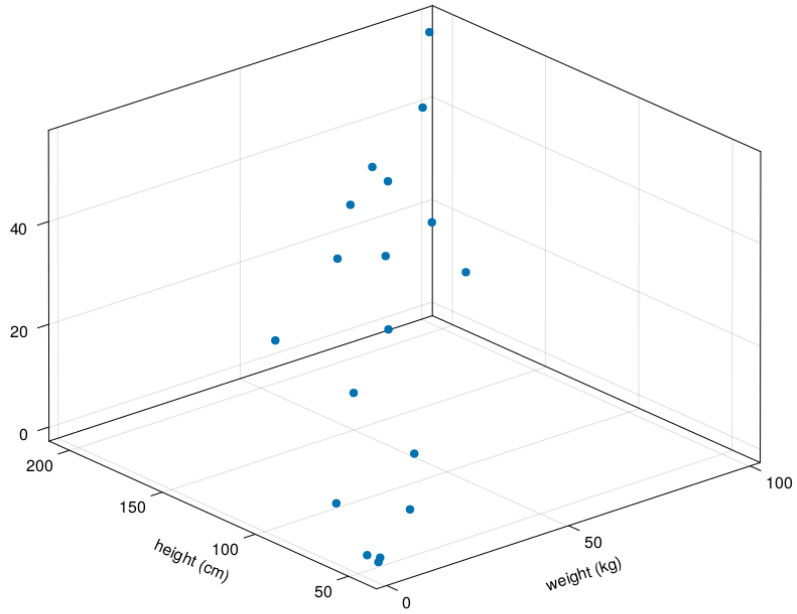


Figure 3: height, weight, and age

What about N > 3?

- No problem!

What about N > 3?

- No problem!
- Unless we want to visualize the data

What about N > 3?

- No problem!
- Unless we want to visualize the data
- Dimensionality reduction

What about N > 3?

- No problem!
- Unless we want to visualize the data
- Dimensionality reduction
  - PCA (Linear Algebra and eigenvalues!)

What about N > 3?

- No problem!
- Unless we want to visualize the data
- Dimensionality reduction
  - PCA (Linear Algebra and eigenvalues!)
  - Self organizing maps

What about non-numeric data?

- I said a vector is a column of numbers, but...

$$\begin{bmatrix} \text{Weight} \\ \text{Height} \\ \text{Systolic blood pressure} \\ \text{Diastolic blood pressure} \\ \text{eGFR} \\ \text{Sex} \\ \text{Race} \\ \text{On ACE inhibitor} \\ \text{On Beta blocker} \end{bmatrix} \tag{2}$$

Section 2

Distance and Similarity

Distances

What are the different ways you can think of to measure "distance"?

"As the crow flies" distance

$$d(\mathbf{x}, \mathbf{x}) = \sqrt{\sum_{\mathbf{l=1}}^{\mathbf{d}} |\mathbf{x_l} - \mathbf{y_l}|^{\mathbf{2}}} \tag{3}$$

- Euclidean distance
- L_2 norm

Manhattan/taxicab distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{l=1}}^{\mathbf{d}} |\mathbf{x_l} - \mathbf{y_l}| \tag{4}$$

- L_1 norm

Maximum distance

- L_∞ norm

$$d(\mathbf{x}, \mathbf{y}) = \mathbf{max_{1 \leq l \leq d}} |\mathbf{x_l} - \mathbf{y_l}| \tag{5}$$

Minkowski distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{\mathbf{l}=\mathbf{1}}^{\mathbf{d}} |\mathbf{x_l} - \mathbf{y_l}|^{\mathbf{p}} \right)^{\frac{\mathbf{1}}{\mathbf{p}}} \tag{6}$$

- previous distances are all special cases of the Minkowski distance

Non-numeric data: Categorical

$$\begin{bmatrix} \text{Sex} \\ \text{Medicare} \\ \text{Retired}|\text{Not retired} \end{bmatrix} \tag{7}$$

Non-numeric data: Binary

$$\delta(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$$

$$d(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{l=1}}^{\mathbf{d}} \delta(\mathbf{x_l}, \mathbf{y_l})$$

- What are potential limitations of this with ordinal data?

Mixed-data types

$$\begin{bmatrix} \text{Weight} \\ \text{Height} \\ \text{Systolic blood pressure} \\ \text{Diastolic blood pressure} \\ \text{eGFR} \\ \text{Sex} \\ \text{Race} \\ \text{On ACE inhibitor} \\ \text{On Beta blocker} \end{bmatrix} \tag{8}$$

Mixed-data types: General distance coefficient

$$d_{gower}(\mathbf{x}, \mathbf{y}) = \left(\frac{\mathbf{1}}{\sum_{\mathbf{l=1}}^{\mathbf{d}} \mathbf{w}(\mathbf{x_l}, \mathbf{y_l})} \sum_{\mathbf{l=1}}^{\mathbf{d}} \mathbf{w}(\mathbf{x_l}, \mathbf{y_l}) \mathbf{d^2}(\mathbf{x_l}, \mathbf{y_l}) \right)^{\frac{\mathbf{1}}{\mathbf{2}}} \tag{9}$$

$w(x_l, y_l)$ is 1 if the comparison is valid

Ordinal data

$$d(x_l, y_l) = \frac{|x_l - y_l|}{R_k}$$

R_k is the range of the kth attribute.

Continuous data

$$d(x_l, y_l) = |x_l - y_l|$$

How would you measure the distance between these data?

- “Brian”
- “Bryan”
- “Brain”
- “Brad”
- “Brianna”
- “Bromine”
- “Broad”

Distance metrics

Formally, a distance metric satisfies the following properties:

- **Non-negativity**: $d(a, b) \geq 0$
- **Identity**: $d(a, a) = 0$
- **Symmetry**: $d(a, b) = d(b, a)$
- **Triangle inequality**: $d(a, c) \leq d(a, b) + d(b, c)$

Distance metrics in Scikit-Learn

https://scikit-learn.org/stable/modules/classes.html#pairwise-metrics

Additional numeric distances ("Chapter 6: Similarity and Dissimilarity Measures," n.d.)

Table 6.1: *Some other dissimilarity measures for numerical data.*

Measure	$d(\mathbf{x}, \mathbf{y})$	Reference
Mean character difference	$\frac{1}{d}\sum_{j=1}^{d}\lvert x_j - y_j\rvert$	Cezkanowski (1909)
Index of association	$\frac{1}{2}\sum_{j=1}^{d}\left\lvert \frac{x_j}{\sum_{l=1}^{d} x_l} - \frac{y_j}{\sum_{l=1}^{d} y_l}\right\rvert$	Whittaker (1952)
Canberra metric	$\sum_{j=1}^{d}\frac{\lvert x_j - y_j\rvert}{(x_j + y_j)}$	Legendre and Legendre (1983)
Czekanowski coefficient	$1 - \frac{2\sum_{j=1}^{d}\min\{x_j, y_j\}}{\sum_{j=1}^{d}(x_j + y_j)}$	Johnson and Wichern (1998)
Coefficient of divergence	$\left(\frac{1}{d}\sum_{j=1}^{d}\left(\frac{x_j - y_j}{x_j + y_j}\right)^2\right)^{\frac{1}{2}}$	Legendre and Legendre (1983)

Figure 4: numeric distances

Additional ordinal distances ("Chapter 6: Similarity and Dissimilarity Measures," n.d.)

Table 6.2: *Some matching coefficients for nominal data.*

Measure	$s(\mathbf{x}, \mathbf{y})$	Weighting of matches, mismatches
Russell and Rao	$\frac{N_{a+d}-N_d}{N_{a+d}+N_{b+c}}$	Equal weights
Simple matching	$\frac{N_{a+d}}{N_{a+d}+N_{b+c}}$	Equal weights
Jaccard	$\frac{N_{a+d}-N_d}{N_{a+d}-N_d+N_{b+c}}$	Equal weights
Unnamed	$\frac{2N_{a+d}}{2N_{a+d}+N_{b+c}}$	Double weight for matched pairs
Dice	$\frac{2N_{a+d}-2N_d}{2N_{a+d}-2N_d+N_{b+c}}$	Double weight for matched pairs
Rogers–Tanimoto	$\frac{N_{a+d}}{N_{a+d}+2N_{b+c}}$	Double weight for unmatched pairs
Unnamed	$\frac{N_{a+d}-N_d}{N_{a+d}-N_d+2N_{b+c}}$	Double weight for unmatched pairs
Kulczynski	$\frac{N_{a+d}-N_d}{N_{b+c}}$	Matched pairs excluded from denominator
Unnamed	$\frac{N_{a+d}}{N_{b+c}}$	Matched pairs excluded from denominator

Figure 5: nominal distances

Additional numeric distances ("Chapter 6: Similarity and Dissimilarity Measures," n.d.)

Table 6.3: *Similarity measures for binary vectors.* $d(\mathbf{x}, \mathbf{y})$ *is the corresponding dissimilarity measure.*

Measure	$s(\mathbf{x}, \mathbf{y})$	Range of $s(\mathbf{x}, \mathbf{y})$	$d(\mathbf{x}, \mathbf{y})$
Jaccard	$\frac{A}{A+B+C}$	$[0, 1]$	$\frac{B+C}{A+B+C}$
Dice	$\frac{A}{2A+B+C}$	$[0, \frac{1}{2}]$	$\frac{B+C}{2A+B+C}$
Pearson	$\frac{AD-BC}{\sigma}$	$[-1, 1]$	$\frac{1}{2} - \frac{AD-BC}{2\sigma}$
Yule	$\frac{AD-BC}{AD+BC}$	$[-1, 1]$	$\frac{BC}{AD+BC}$
Russell–Rao	$\frac{A}{d}$	$[0, 1]$	$1 - \frac{A}{d}$
Sokal–Michener	$\frac{A+D}{d}$	$[0, 1]$	$\frac{2(B+C)}{A+2(B+C)+D}$
Rogers–Tanimoto	$\frac{A+D}{A+2(B+C)+D}$	$[0, 1]$	$\frac{2(B+C)}{A+2(B+C)+D}$
Rogers–Tanimoto-a	$\frac{A+D}{A+2(B+C)+D}$	$[0, 1]$	$\frac{2(d-A-D)}{2d-A-D}$
Kulzinsky	$\frac{A}{B+C}$	$[0, \infty]$	$\frac{B+C-A+d}{B+C+d}$

Figure 6: binary distances

Section 3

Cluster quality

Cluster quality

How does the scatter/spread within a cluster compare to scatter/spread across clusters?

https://scikit-learn.org/stable/modules/classes.html#clustering-metrics

Section 4

Algorithms

Agglomerative hierarchical clustering

- Assign each item to a unique cluster (N data elements $\rightarrow$ N clusters)
  - Identify two closest clusters and merge
  - Repeat until only single cluster

Linkage Metrics: how to identify closest clusters

- *Single-linkage*: minimum point-to-point distance
- *Complete-linkage*: maximum point-to-point distance
- *Average-linkage*: average point-to-point distance

Hierarchical Clustering: Example[1]

	BOS	NYC	CHI	DEN	SF	SEA
BOS	0	206	963	1949	3095	2970
NYC		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0

[1] Courseware (2016)

Hierarchical Clustering single-linkage: Step 1

	BOS	NYC	CHI	DEN	SF	SEA
BOS	0	206	963	1949	3095	2970
NYC		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0

`{BOS} {NYC} {CHI} {DEN} {SF} {SEA}`

Hierarchical Clustering single-linkage: Step 2

	BOS	NYC	CHI	DEN	SF	SEA
BOS	0	206	963	1949	3095	2970
NYC		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0

```
{BOS} {NYC} {CHI} {DEN} {SF} {SEA}
{BOS, NYC} {CHI}  {DEN} {SF} {SEA}
```

Hierarchical Clustering single-linkage: Step 3

	BOS	NYC	CHI	DEN	SF	SEA
BOS	0	206	963	1949	3095	2970
NYC		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0

```
{BOS} {NYC} {CHI} {DEN} {SF} {SEA}
{BOS, NYC} {CHI}  {DEN} {SF} {SEA}
{BOS, NYC, CHI} {DEN} {SF} {SEA}
```

Hierarchical Clustering single-linkage: Step 4

	BOS	NYC	CHI	DEN	SF	SEA
BOS	0	206	963	1949	3095	2970
NYC		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0

```
{BOS} {NYC} {CHI} {DEN} {SF} {SEA}
{BOS, NYC} {CHI}  {DEN} {SF} {SEA}
{BOS, NYC, CHI} {DEN} {SF} {SEA}
{BOS, NYC, CHI} {DEN} {SF, SEA}
```

Hierarchical Clustering single-linkage: Step 5

	BOS	NYC	CHI	DEN	SF	SEA
BOS	0	206	963	1949	3095	2970
NYC		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0

```
{BOS} {NYC} {CHI} {DEN} {SF} {SEA}
{BOS, NYC} {CHI}  {DEN} {SF} {SEA}
{BOS, NYC, CHI} {DEN} {SF} {SEA}
{BOS, NYC, CHI} {DEN} {SF, SEA}
{BOS, NYC, CHI, DEN} {SF, SEA}
```

Hierarchical Clustering single-linkage: Step 6

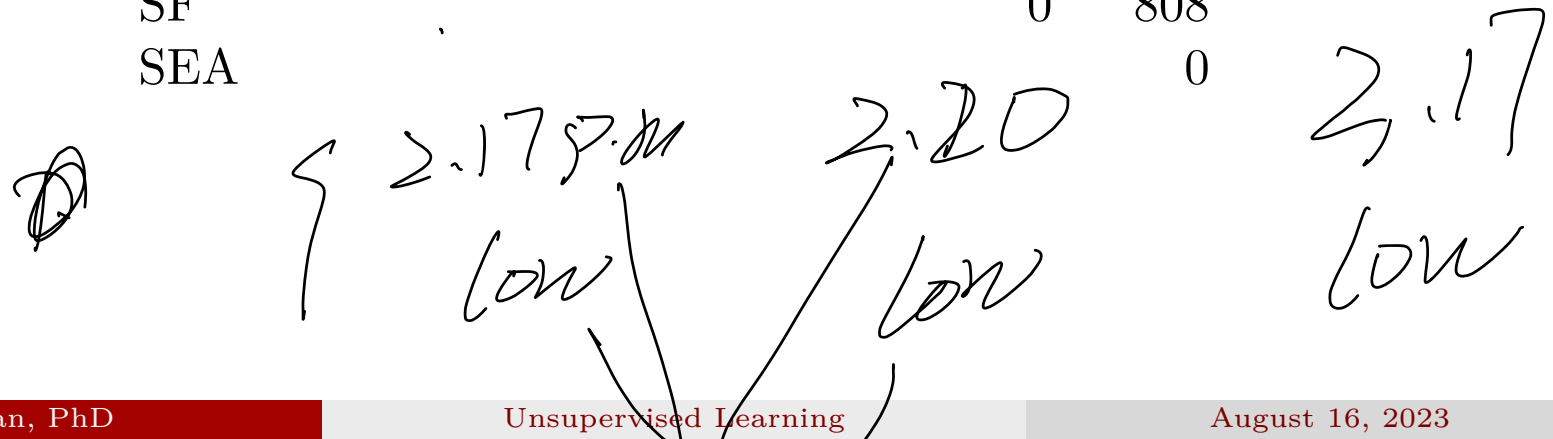
	BOS	NYC	CHI	DEN	SF	SEA
BOS	0	206	963	1949	3095	2970
NYC		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0

```
{BOS} {NYC} {CHI} {DEN} {SF} {SEA}
{BOS, NYC} {CHI}  {DEN} {SF} {SEA}
{BOS, NYC, CHI} {DEN} {SF} {SEA}
{BOS, NYC, CHI} {DEN} {SF, SEA}
{BOS, NYC, CHI, DEN, SF, SEA}
```

Hierarchical Clustering with total-linkage?

{SEA}

	BOS	NYC	CHI	DEN	SF	SEA
BOS	0	206	963	1949	3095	2970
NYC		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0



Hierarchical: Good and Bad

- see the history
- deterministic
  - given a linkage, always get the same result
- See whole history
- Greedy algorithm
  - locally optimal may not be globally optimal
- Potentially really, really slow ($O(n^3)$)

K-Means

```
randomly chose k items as initial centroids

while True:
    create k clusters from centroids and data
       assign each item to closest cedntroid
    compute k new centroids
       cluster averages
    if centroids identical to previous iteration:
        break
```

K-Means

```
randomly chose k items as initial centroids

while True:
    create k clusters
       assign each item to closest cedntroid
    compute k new centroids
       cluster averages
    if centroids identical to previous iteration:
        break
```

Is this algorithm deterministic?

K-Means Example

The following screenshots are from “Introduction to Computational Thinking and Data Science.”(Courseware 2016)

k-means Example

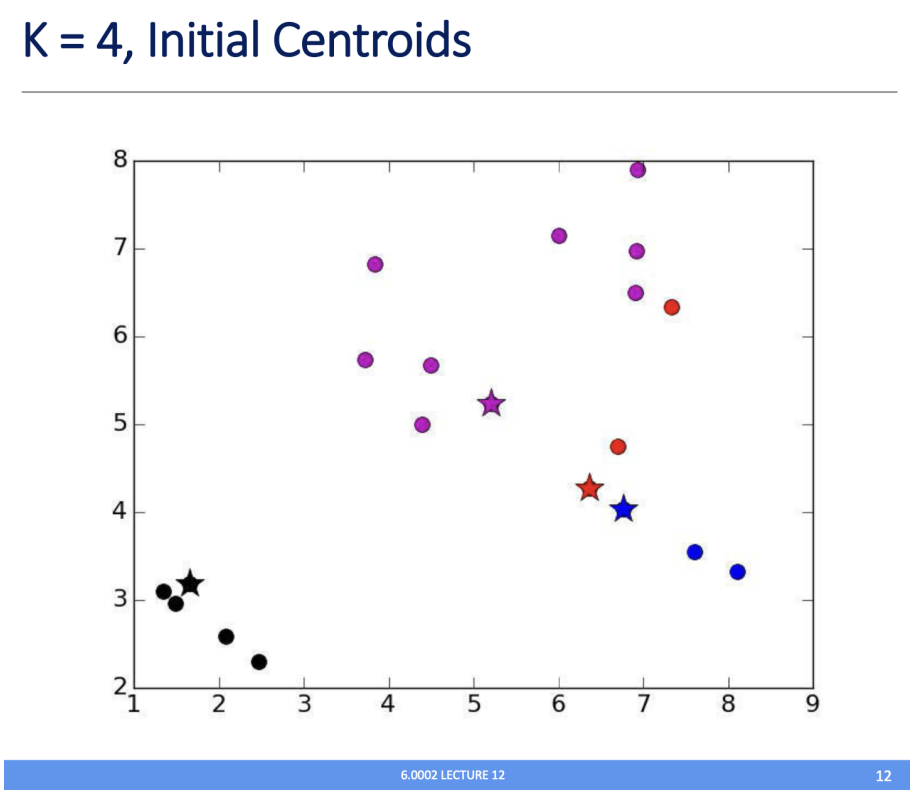


Figure 7: initial centroids

k-means Example

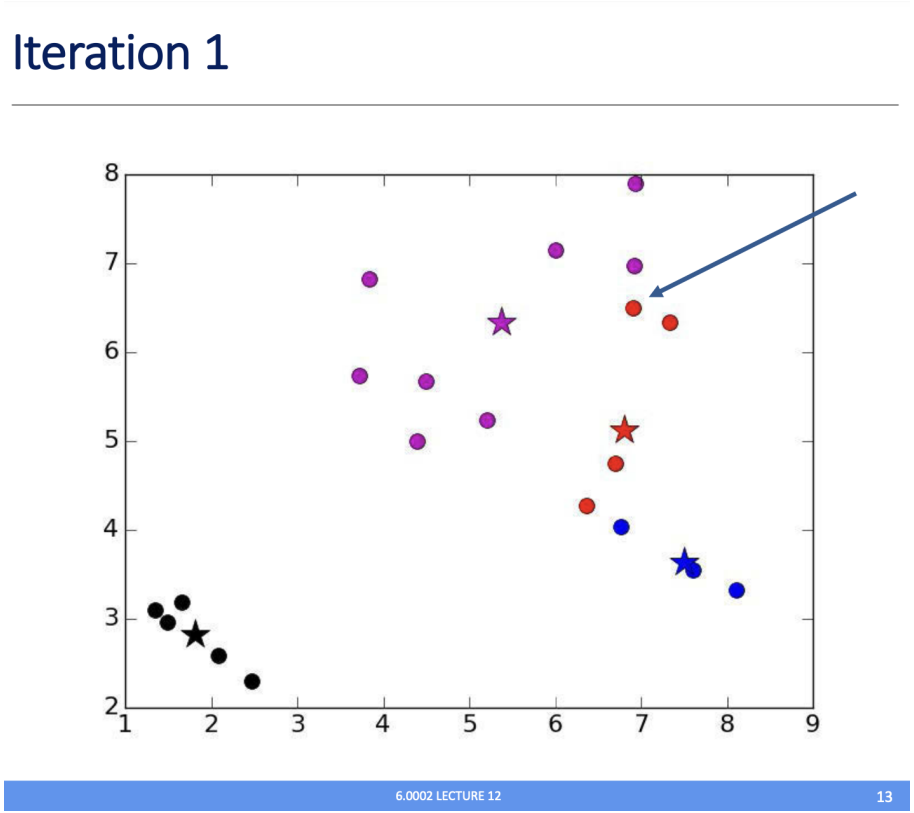


Figure 8: initial centroids

k-means Example

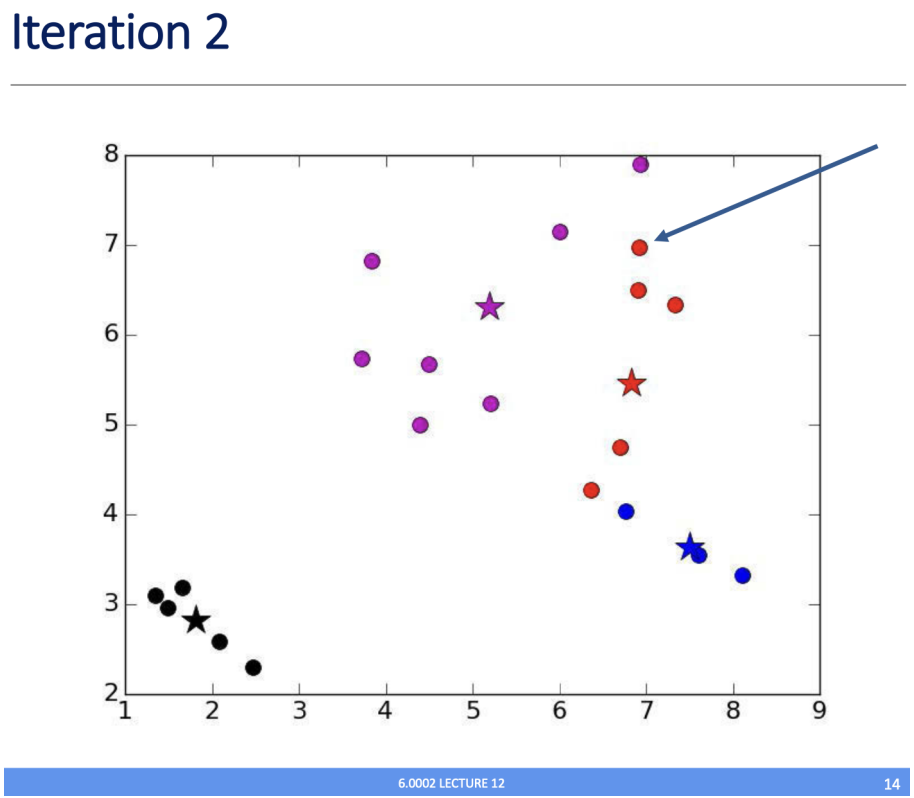


Figure 9: initial centroids

k-means Example

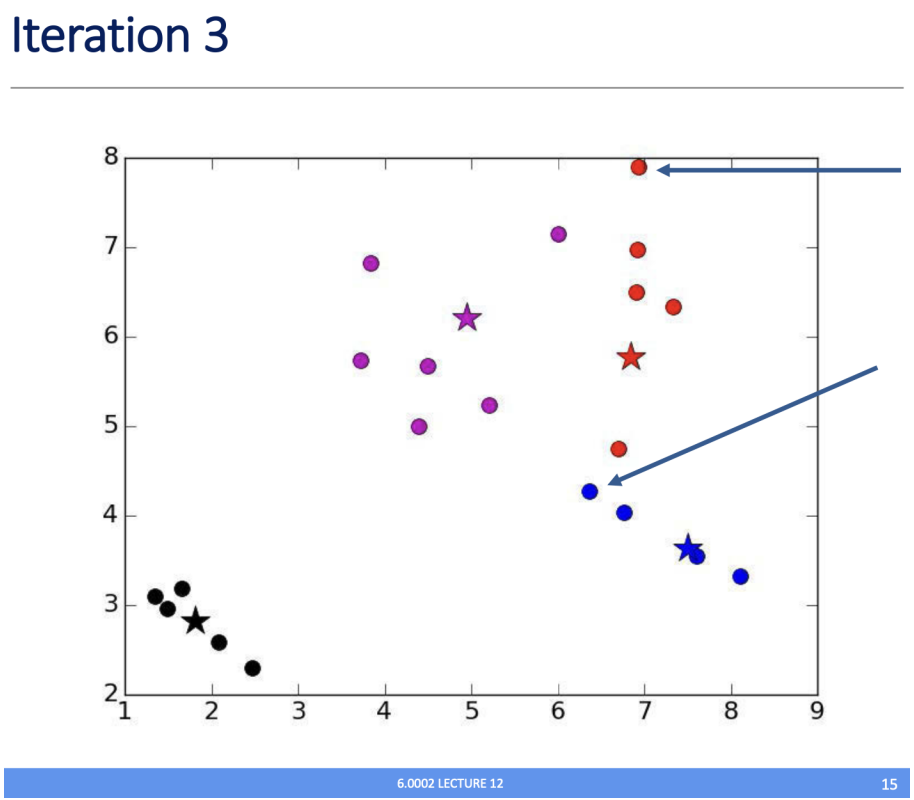


Figure 10: initial centroids

k-means Example

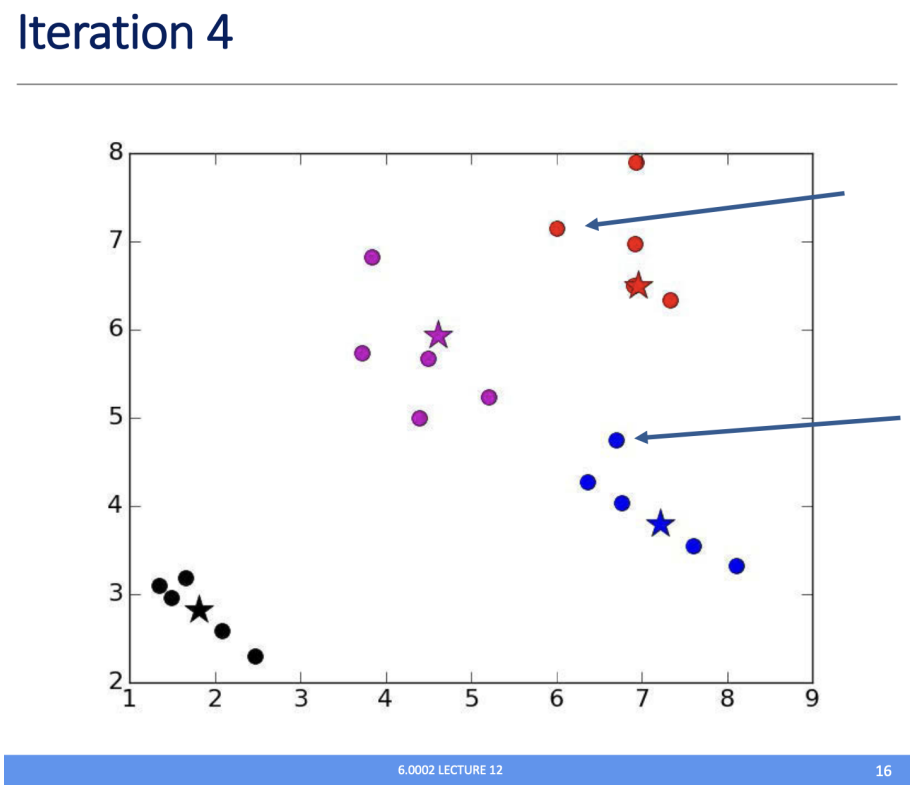


Figure 11: initial centroids

k-means Example

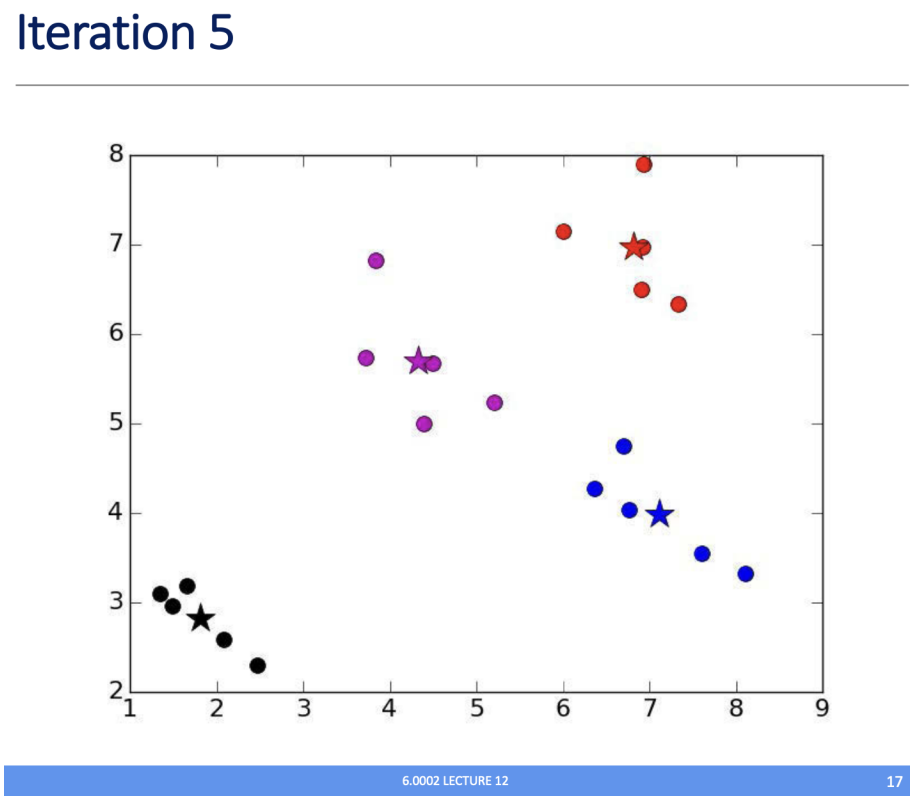


Figure 12: initial centroids

Example: Iris dataset clustering

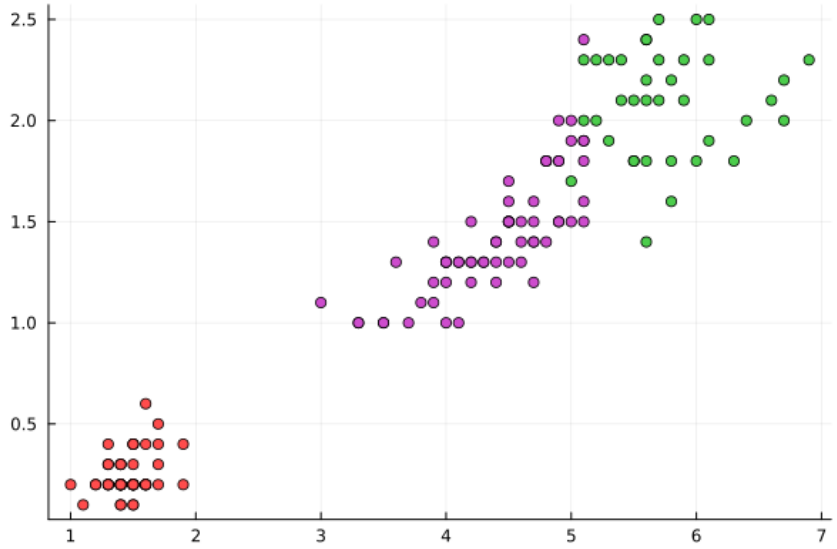


Figure 13: good

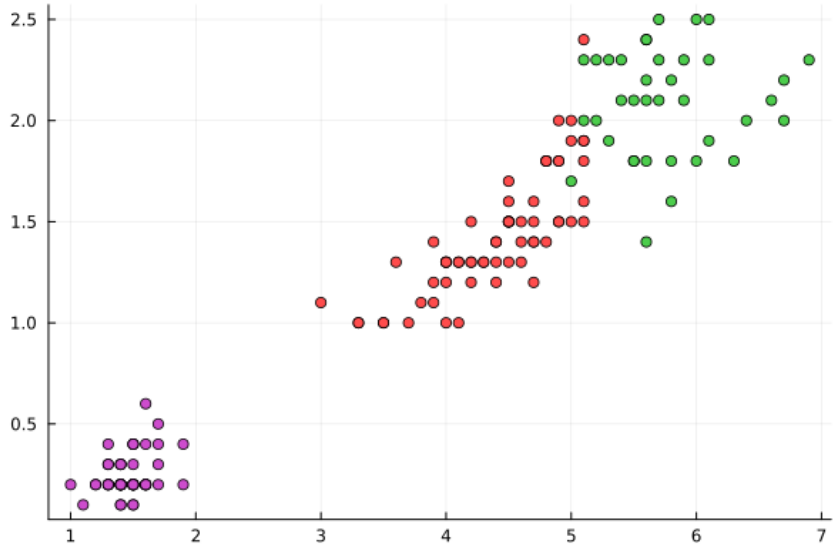


Figure 14: bad

k-means: stochastic search

- Run multiple times with different seeds
- Select run with best clustering metric

k-means: smart selection of seeds

- "k-means++: the advantages of careful seeding" (Arthur and Vassilvitskii 2007)
- "A simple and fast algorithm for K-medoids clustering" (Park and Jun 2009)

Selection of k

- The number of means (k) is the constraint on the optimization problem
- How to choose k?
  - domain knowledge
    - Gray matter, white matter, cerebral spinal fluid
    - five families of bacteria
  - hierarchical clustering
    - random samples of points
  - Plot of cluster dissimilarity vs number of clusters (“elbow plot”)

Example: Elbow plot iris data set

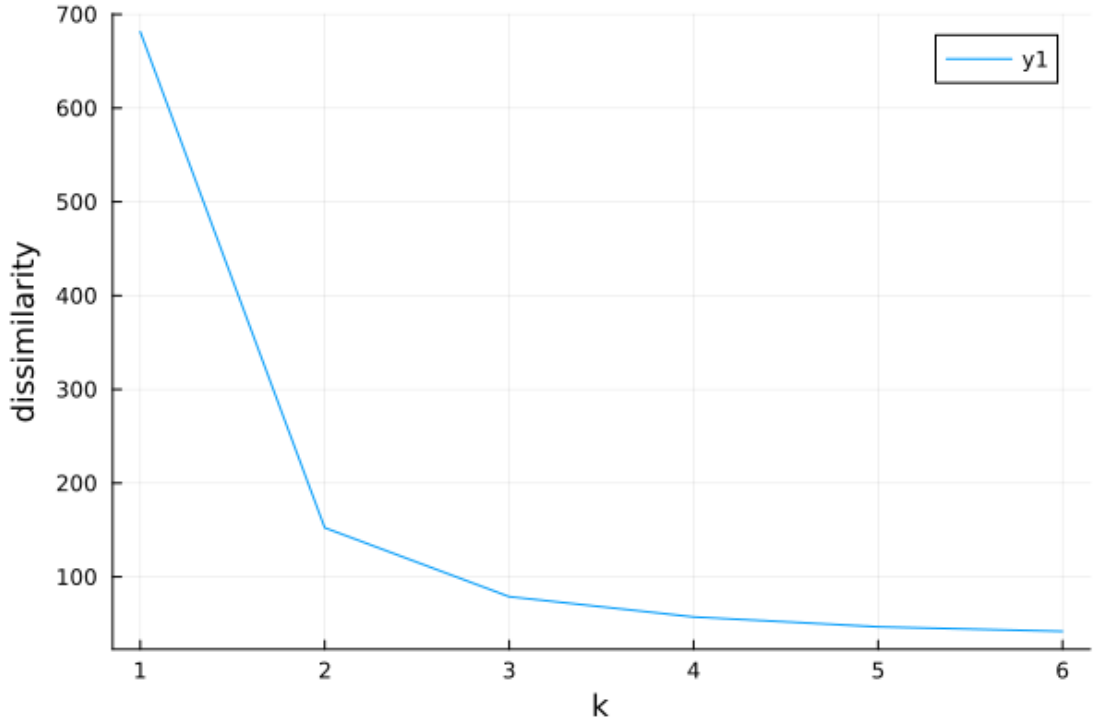


Figure 15: iris elbow plot

Main ingredients for clustering

- Similarity/dissimilarity metric
- A function to evaluate cluster quality
  - Intra-cluster cohesion
  - Inter-cluster separation
  - Clustering is subjective
- Clustering algorithm
  - That optimizes an evaluation function
- How many clusters?
  - Fixed k clusters
  - Find the best k to optimize a function

k-means

- Visualizing K-Means Clustering
- Complexity $O(N^2)$ or $O(N \log N)$

DBSCAN

"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" (Ester et al. 1996)

- Discover the number of clusters

DBSCAN

"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" (Ester et al. 1996)

- Discover the number of clusters
- What constraints to provide?

DBSCAN

"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" (Ester et al. 1996)

- Discover the number of clusters
- What constraints to provide?
  - Minimum number of elements in a cluster

DBSCAN

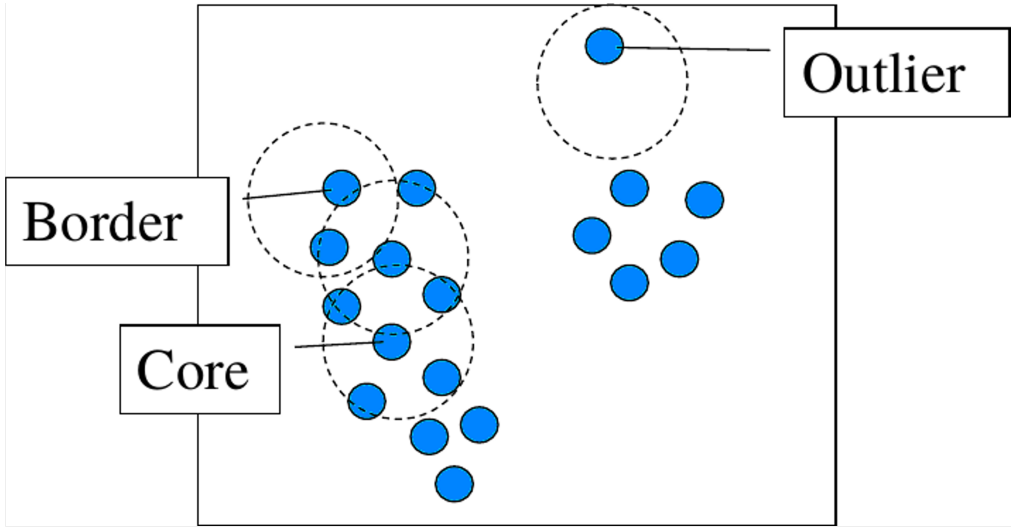
"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" (Ester et al. 1996)

- Discover the number of clusters
- What constraints to provide?
  - Minimum number of elements in a cluster
  - Maximum distance between neighbors

DBSCAN

```
1:  procedure DBSCAN(X, eps, minpts)
2:      for each unvisited point x ∈ X do
3:          mark x as visited
4:          N ← GETNEIGHBORS(x, eps)
5:          if |N| < minpts then
6:              mark x as noise
7:          else
8:              C ← {x}
9:              for each point x' ∈ N do
10:                 N ← N \ x'
11:                 if x' is not visited then
12:                     mark x' as visited
13:                     N' ← GETNEIGHBORS(x', eps)
14:                     if |N'| ≥ minpts then
15:                         N ← N ∪ N'
16:                 if x' is not yet member of any cluster then
17:                     C ← C ∪ {x'}
```

Figure 16: algorithm



ε =1

#minPts = 5

Figure 17: example

DBSCAN

- Visualizing DBSCAN Clustering

References I

Arthur, David, and Sergei Vassilvitskii. 2007. "K-Means++: The Advantages of Careful Seeding." In *SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–35. Philadelphia, PA, USA: Society for Industrial; Applied Mathematics.

"Chapter 6: Similarity and Dissimilarity Measures." n.d. In *Data Clustering: Theory, Algorithms, and Applications, Second Edition*, 65–100. https://doi.org/10.1137/1.9781611976335.ch6.

Courseware, MIT Open. 2016. "Introduction to Computational Thinking and Data Science." https://ocw.mit.edu/courses/6-0002-introduction-to-computational-thinking-and-data-science-fall-2016/.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–31. KDD'96. Portland, Oregon: AAAI Press.

References II

Park, Hae-Sang, and Chi-Hyuck Jun. 2009. "A Simple and Fast Algorithm for k-Medoids Clustering." *Expert Systems with Applications* 36 (2, Part 2): 3336–41. https://doi.org/https://doi.org/10.1016/j.eswa.2008.01.039.

data fetch

一、药物 ⇒ subject_id, hadm_id Cdata①) ⇒ 药物⇒item item_id. data④*

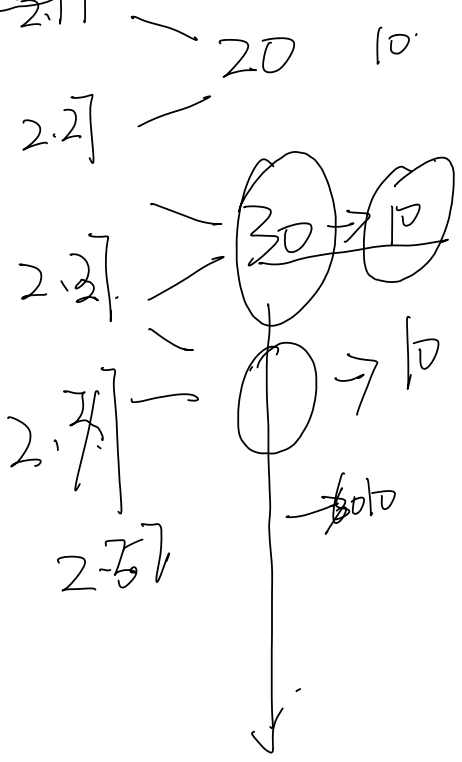
二、血压指标 ⇒ d_item itemid for BP Cdata③)

三、data① subject_id ⇒ chartevents (subject_id, hadm_id, charttime, valuenum, 单位) data③
data② 血压

四、data① subject_id ⇒ inputevents (subject_id, hadm_id, starttime, endtime, itemid) data④
data①* item_id

在对research查了相关药物
definition for BP for hopo tension profound

Assumption 给药后的首次血压测量即被作
低⇒即出现低血压(hypotension)



diff [value in BP] normal

diff(charttime) 0

for a patient 2:17 pm

2:27 pm

2:37 pm

2:47 pm

3:47 pm

low → 10 min

low 10 min

low 10 min

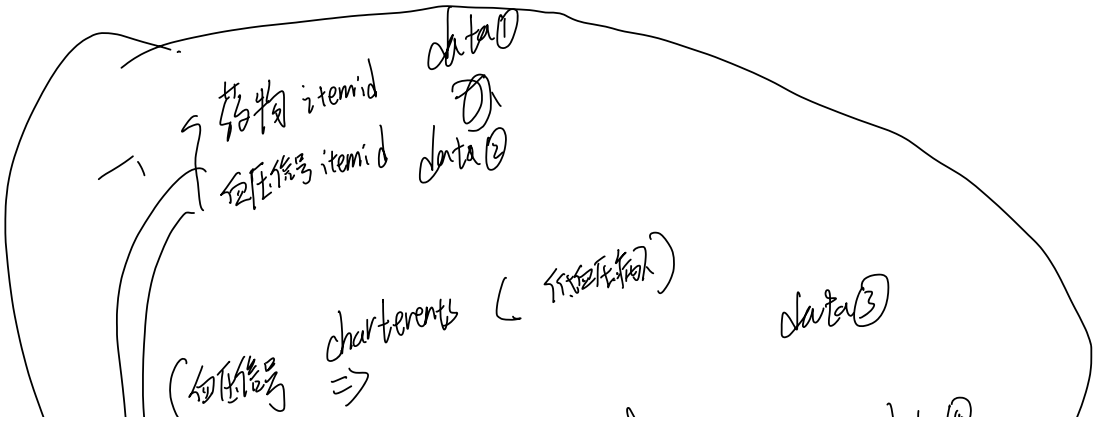
10 60 min

10

同时间换用标签

to simplify

若 血压标签显示为 low [itemid in [ ]] => 就是hypotension



data④

三、药物名称 => inputevents 低血压用药病人

subject, hadm, stay from data③

两次测量血压值差 < -60 且 测量时间间隔超过 60min 且 血压值的标签里有低血压信号.

三条件在chart table中同时满足 => 认定他满足 pround hypotension 的症状

在 med-data中 若病人的服药开始时间是在测量之后，且结束之后 测血压信号 [illegible] hypotension => 则认定他是 profound hypotension

每个病人每次住院都有不一样的病情

hadm_id =>

加一个超链接 => itemid for Blood pressure

大图展示所有的 Blood pressure 然后标记哪些升压药的手工用到的.

未做

① 把以前的过滤的列 加回来 [ditt[tim], ditt[value]]，打的结

② 加 True False for ICD_code

但由最

选的一组

First step

合并

[用现在的表 合并后的] ⇐ ditt[time] ditt[value]

Second step

[ICD 码表里的 hadm_id, subject_id 对的 True/False]

通过观察总结数据 ⇒ 可以有是一个好的可接受的区理

此外，通过浏览两个网站发现

ICD 更加 generic 是一个通用的代码。

SMOTE 更加的全面，对于处理的方法更加详细

(比如：……)