Ch 12.1, 12.4: Unsupervised Learning & Clustering

Lecture 32 - CMSE 381

Prof. Lianzhang Bao

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Mon, Apr 14, 2025

Announcements

Last time:

- Convolutional Neural Nets

This lecture:

- Clustering (Just hierarchical clustering)

Announcements:

- No more homework!
- Weds: Project office hours, zoom only!
- Fri Apr 18 : Review - Bring questions!
- Mon Apr 21: Exam
  - Content since 2nd Exam (Ch 7 and on)
  - One page (8.5x11) handwritten cheat sheet
  - Calculator if you want it

21	W	3/19	Polynomial & Step Functions	7.1-7.2		
22	F	3/21	Step Functions; Basis functions; Start Splines	7.2-7.4		
23	M	3/24	Regression Splines	7.4		
24	W	3/26	Decision Trees	8.1	HW #6 Due Wed 3/26	Q7
25	F	3/28	Random Forests	8.2.1, 8.2.2	HW #7 Due Sun 3/30	
26	M	3/31	Maximal Margin Classifier	9.1		
27	W	4/2	SVC	9.2		Q8
28	F	4/4	SVM	9.3, 9.4	HW #8 Due Sun 4/6	
29	M	4/7	Single Layer NN	10.1		
30	W	4/9	Multi Layer NN	10.2		Q9
31	F	4/11	CNN	10.3	HW #9 Due Sun 4/13	
32	M	4/14	Unsupervised learning / clustering	12.1, 12.4		
33	W	4/16	Virtual: Project Office Hours			Q10
	F	4/18	***Review***			
	M	4/21	**Midterm #3**			
	W	4/23				
	F	4/25			**Project Due**	
			No final exam			

Section 1

Unsupervised learning

Supervised vs Unsupervised Learning

Supervised

Unupervised

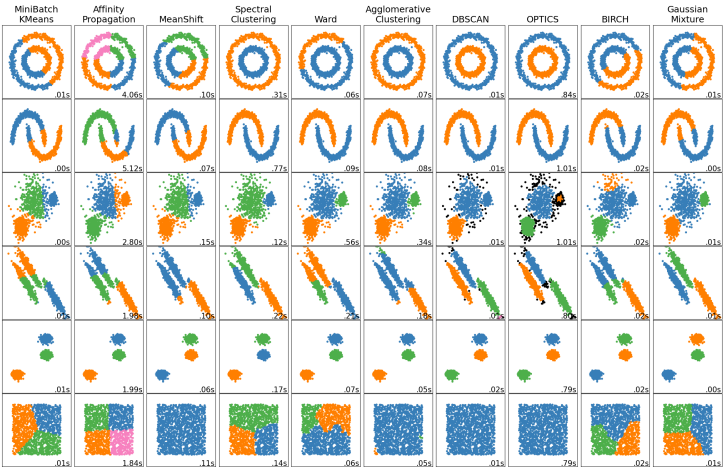
Some examples of unsupervised problems

- Assay gene expression levels in 100 patients with breast cancer, looking for subgroups with similar qualities
- Online shopping: find groups of shoppers with similar browsing and purchase histories and show relevant related products.
- Search engine picking results to show

Section 2

Clustering

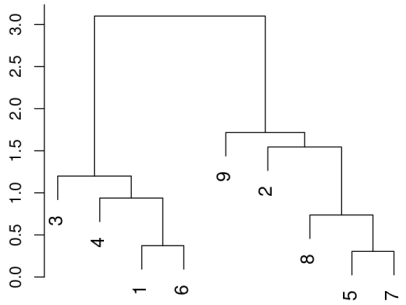
Big idea

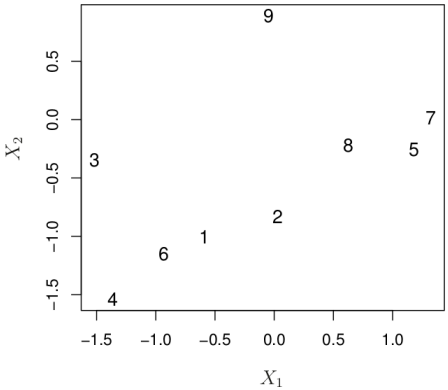

MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
OPTICS
BIRCH
Gaussian Mixture
.01s
4.06s
.10s
.31s
.06s
.07s
.01s
.84s
.02s
.01s
.00s
5.12s
.07s
.77s
.09s
.08s
.01s
1.01s
.02s
.00s
.00s
2.80s
.15s
.12s
.56s
.34s
.01s
1.01s
.02s
.01s
.01s
1.98s
.10s
.22s
.22s
.19s
.02s
.80s
.02s
.01s
.01s
1.99s
.06s
.17s
.07s
.05s
.02s
.79s
.02s
.00s
.01s
1.84s
.11s
.14s
.06s
.05s
.01s
.79s
.02s
.01s

Section 3

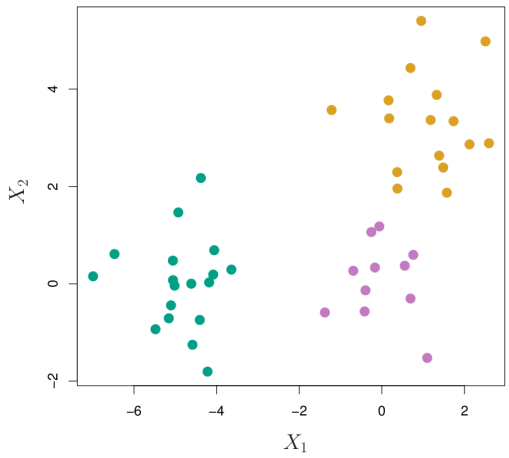
Hierarchical Clustering

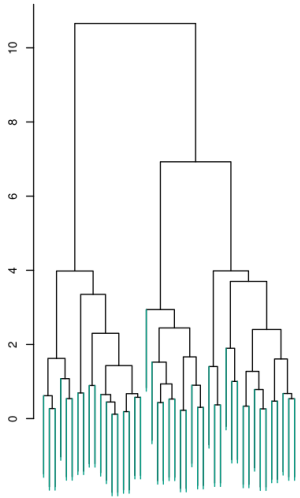
Dendrogram


0.0
0.5
1.0
1.5
2.0
2.5
3.0
3
4
1
6
9
2
8
5
7


9
7
3
8
5
2
1
6
4
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
X_2
X_1

A bigger example

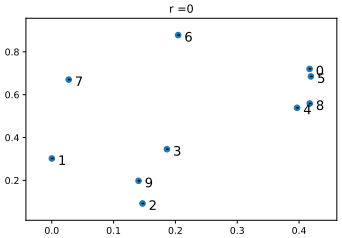


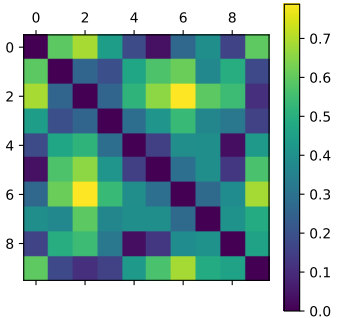


Single linkage

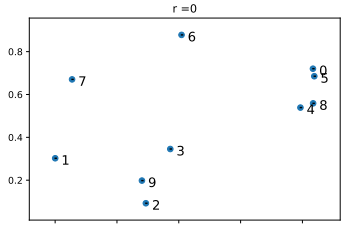
Distance between cluster A and cluster B:
Smallest distance between the points

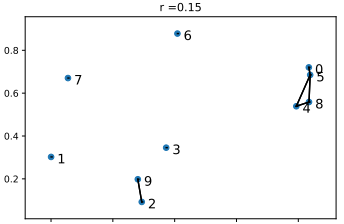
$$L(A, B) = \min_{a \in A, b \in B} \|a - b\|$$

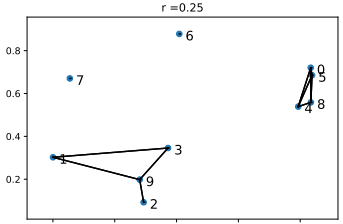


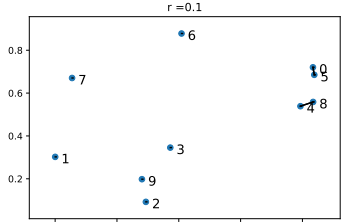


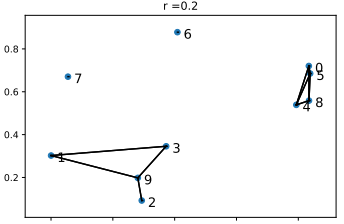
Building the dendrogram

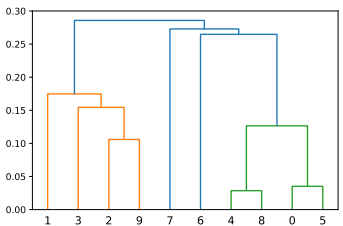




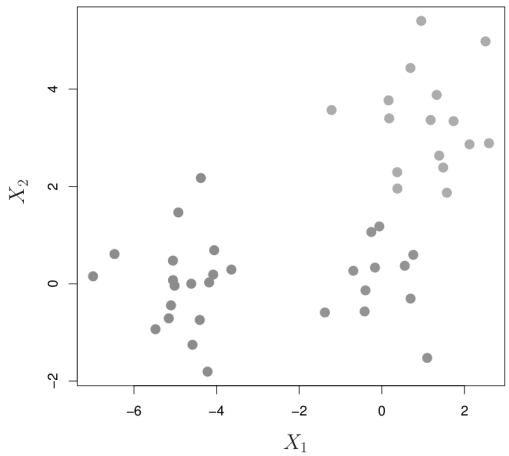


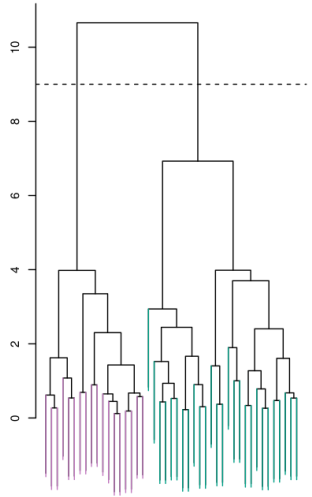




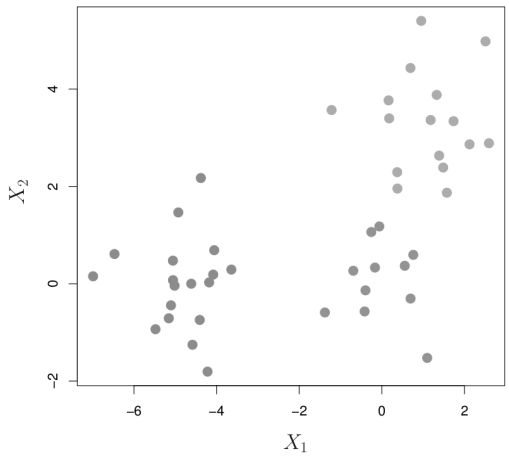


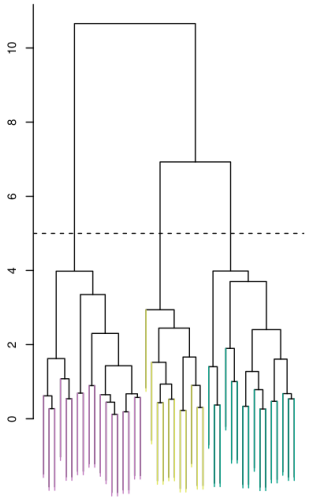
How to get clusters



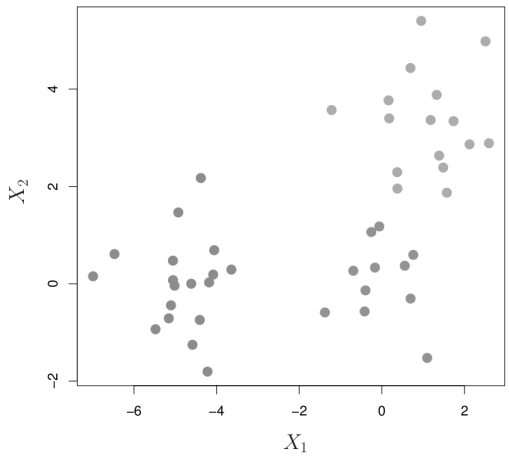


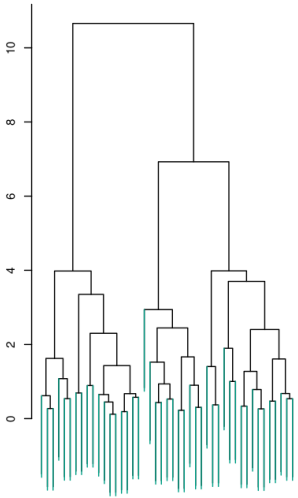
How to get different clusters





Can get any number of clusters

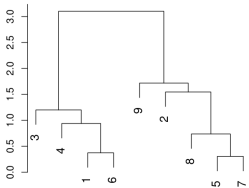




Linkage

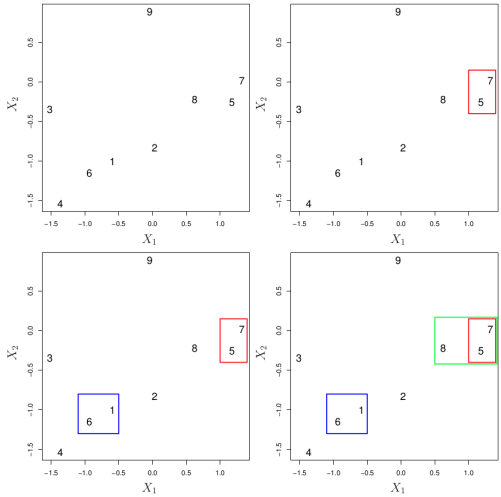
Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

Example with complete linkage

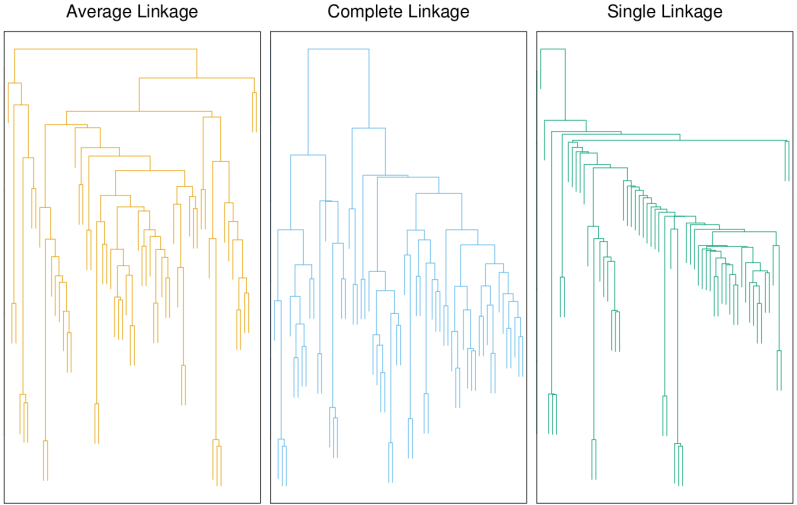


Distance between cluster A and cluster B:
Largest distance between the points

$$L(A, B) = \max_{a \in A, b \in B} \|a - b\|$$



Examples of different linkage



Dependence on dissimilarity measure

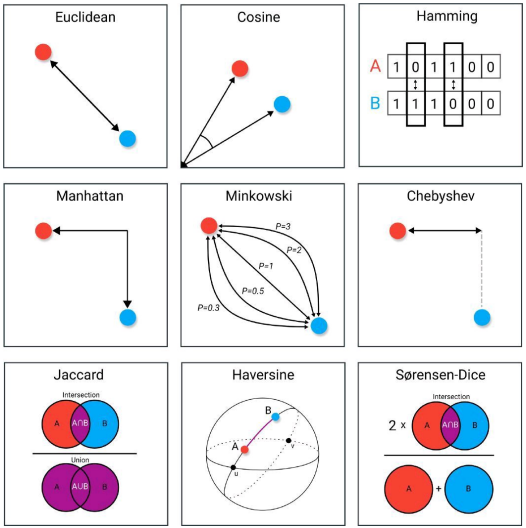


Photo Credit Link

Coding

Next time

21	W	3/19	Polynomial & Step Functions	7.1-7.2		
22	F	3/21	Step Functions; Basis functions; Start Splines	7.2-7.4		
23	M	3/24	Regression Splines	7.4		
24	W	3/26	Decision Trees	8.1	HW #6 Due Wed 3/26	Q7
25	F	3/28	Random Forests	8.2.1, 8.2.2	HW #7 Due Sun 3/30	
26	M	3/31	Maximal Margin Classifier	9.1		
27	W	4/2	SVC	9.2		Q8
28	F	4/4	SVM	9.3, 9.4	HW #8 Due Sun 4/6	
29	M	4/7	Single Layer NN	10.1		
30	W	4/9	Multi Layer NN	10.2		Q9
31	F	4/11	CNN	10.3		
32	M	4/14	Unsupervised learning / clustering	12.1, 12.4	HW #9 Due Sun 4/13	
33	W	4/16	Virtual: Project Office Hours			Q10
	F	4/18	***Review***			
	M	4/21	**Midterm #3**			
	W	4/23				
	F	4/25			**Project Due**	
			No final exam			