Ch 12.1, 12.4: Unsupervised Learning & Clustering

Lecture 32 - CMSE 381

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

Mon, Apr 13, 2026

Announcements

Last time:

- Convolutional Neural Nets

This lecture:

- Clustering (Just hierarchical clustering)

Announcements:

- Fri 4/17: Review - submit your questions here!
- Monday 4/20: Exam 3
  - Content since 2nd Exam (Ch 7 and on)
  - One page (8.5x11) handwritten cheat sheet
  - no-internet Calculator

21	W	3/18	Polynomial & Step Functions	7.1-7.2		
22	F	3/20	Step Functions; Basis functions; Start Splines	7.2-7.4		
23	M	3/23	Regression Splines	7.4		
24	W	3/25	Decision Trees	8.1		Q7
25	F	3/27	Random Forests	8.2.1, 8.2.2	HW #5 Due Sun 3/29	
26	M	3/30	Maximal Margin Classifier	9.1		
27	W	4/1	SVC	9.2		Q8
28	F	4/3	SVM	9.3, 9.4		
29	M	4/6	Single Layer NN	10.1		
30	W	4/8	Multi Layer NN	10.2		Q9
31	F	4/10	CNN	10.3		
32	M	4/13	Unsupervised learning / clustering	12.1, 12.4	HW #6 Due Sun 4/12	
33	W	4/15	Virtual: Project Office Hours			Q10
	F	4/17	***Review***			
	M	4/20	**Midterm #3**			
	W	4/22				
	F	4/24			**Project Due**	

What will you learn today?

- What is the difference between supervised vs unsupervised learning?
- What do clustering methods aim to accomplish?
- How to interpret a dendrogram of hierarchical clustering?
- How are different linkage methods defined?
- How to perform hierarchical clustering in Python?

Section 1

Unsupervised learning

Supervised vs Unsupervised Learning

Supervised

Unsupervised

Some examples of unsupervised problems

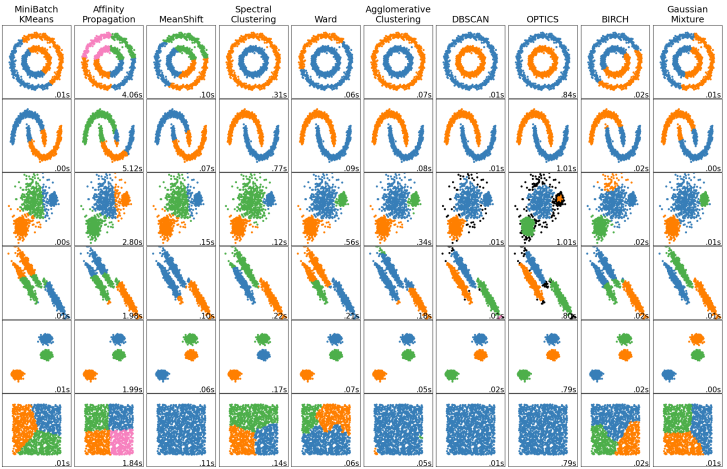
- Assay gene expression levels in 100 patients with breast cancer, looking for subgroups with similar qualities
- Online shopping: find groups of shoppers with similar browsing and purchase histories and show relevant related products.
- Search engine picking results to show

Section 2

Clustering

Big idea

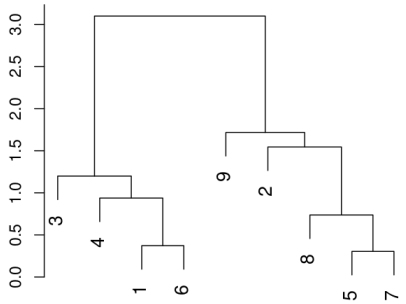
Clustering: relation between samples

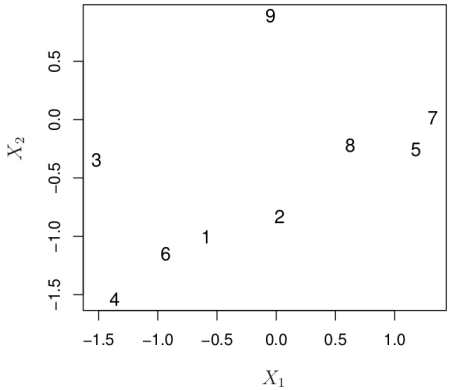


Section 3

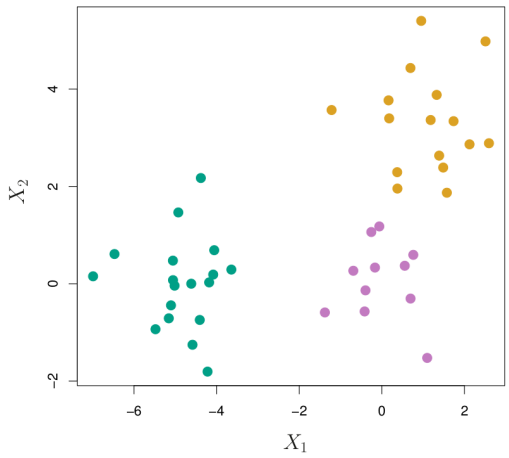
Hierarchical Clustering

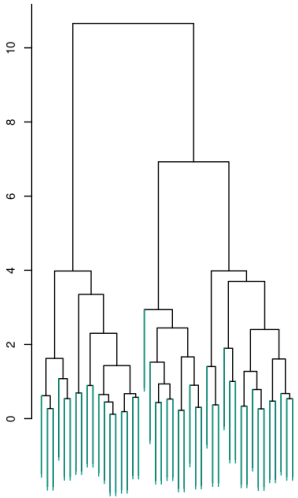
Dendrogram





A bigger example

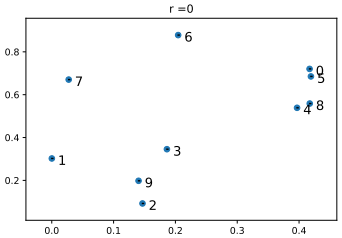


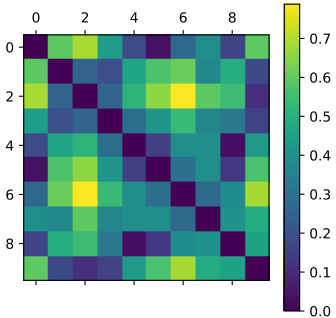


Single linkage

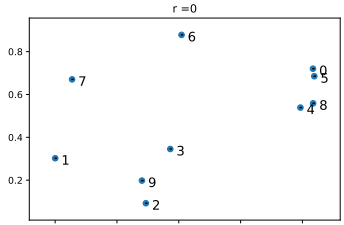
Distance between cluster A and cluster B:
Smallest distance between the points

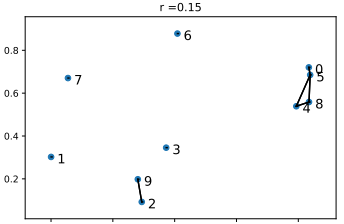
$$L(A, B) = \min_{a \in A, b \in B} \|a - b\|$$

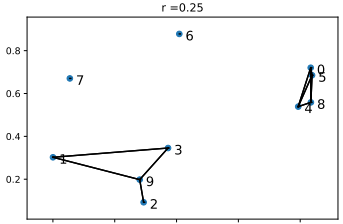


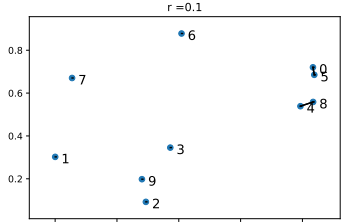


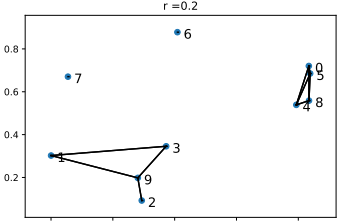
Building the dendrogram

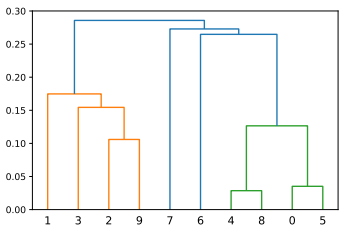




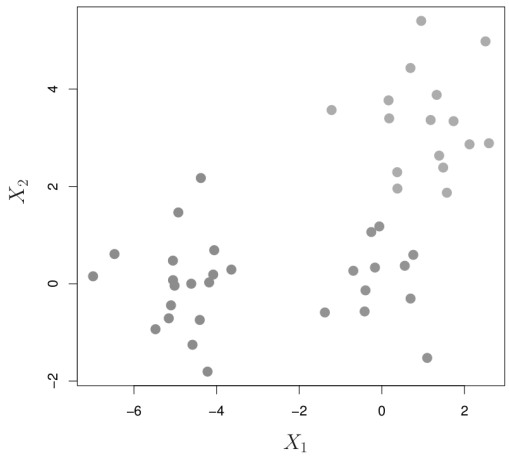


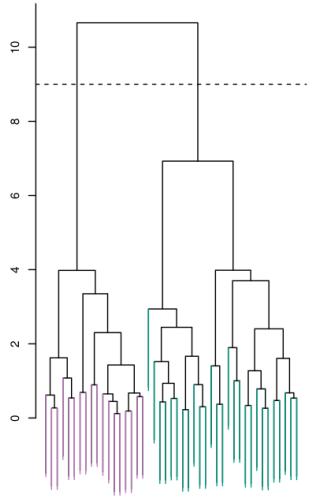




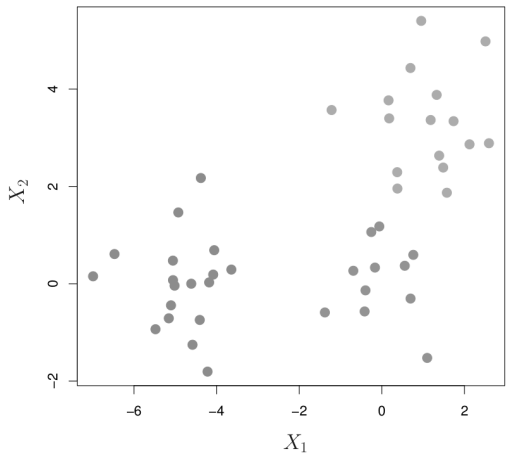


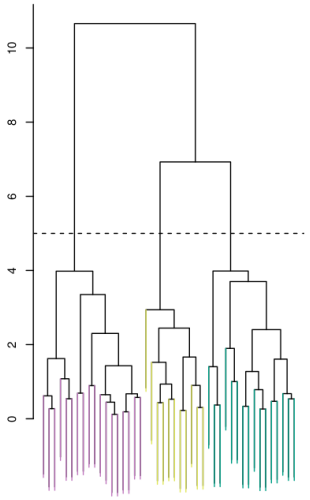
How to get clusters



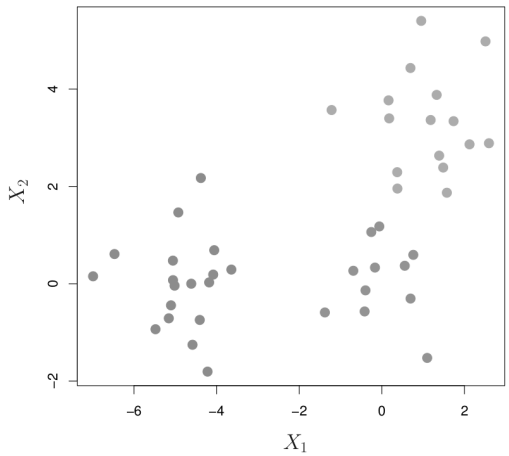


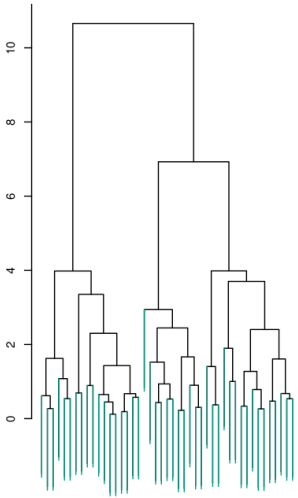
How to get different clusters





Can get any number of clusters



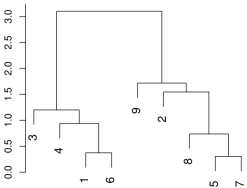


Test your understanding: PollEv

Linkage

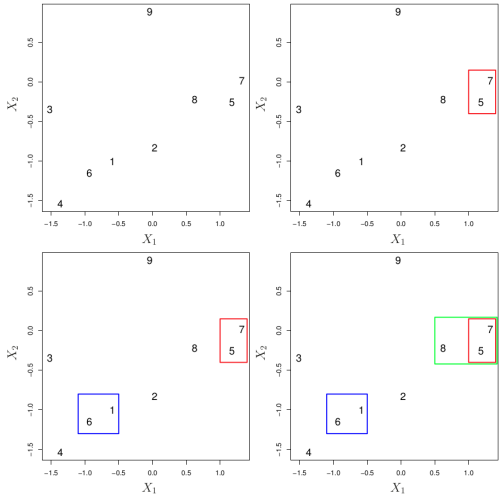
Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities 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

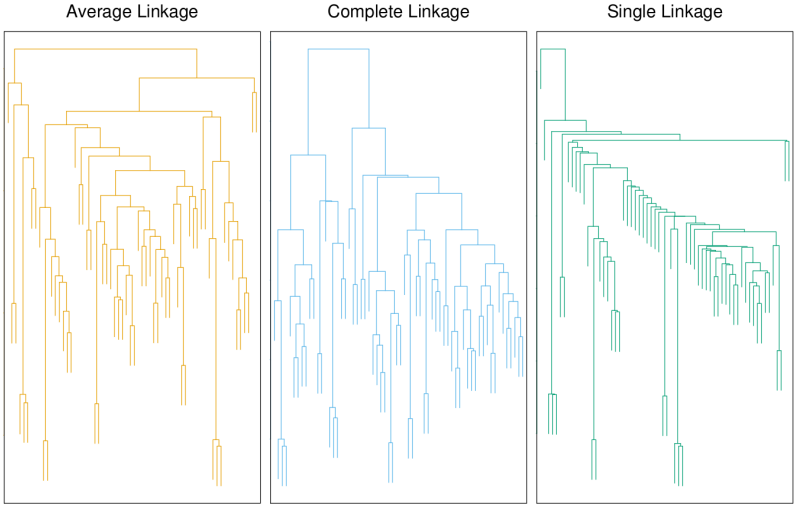


Distance between cluster A and cluster B:
Largest distance between the points

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



Examples of different linkage



Dependence on dissimilarity measure

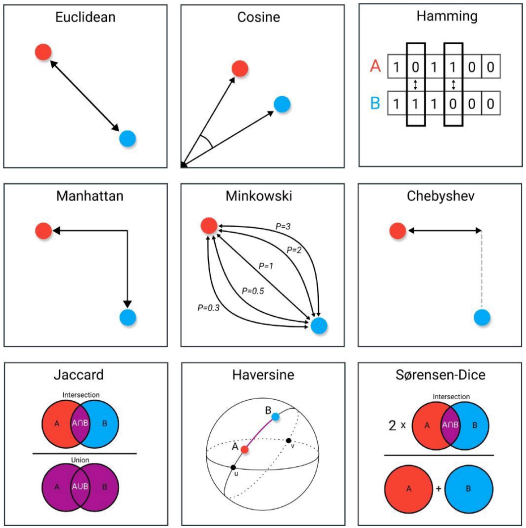


Photo Credit Link

Coding

Next time

21	W	3/18	Polynomial & Step Functions	7.1-7.2		
22	F	3/20	Step Functions; Basis functions; Start Splines	7.2-7.4		
23	M	3/23	Regression Splines	7.4		
24	W	3/25	Decision Trees	8.1		Q7
25	F	3/27	Random Forests	8.2.1, 8.2.2	HW #5 Due Sun 3/29	
26	M	3/30	Maximal Margin Classifier	9.1		
27	W	4/1	SVC	9.2		Q8
28	F	4/3	SVM	9.3, 9.4		
29	M	4/6	Single Layer NN	10.1		
30	W	4/8	Multi Layer NN	10.2		Q9
31	F	4/10	CNN	10.3		
32	M	4/13	Unsupervised learning / clustering	12.1, 12.4	HW #6 Due Sun 4/12	Q10
33	W	4/15	Virtual: Project Office Hours			
	F	4/17	***Review***			
	M	4/20	**Midterm #3**			
	W	4/22				
	F	4/24			**Proiect Due**	