Ch 12.1, 12.4: Unsupervised Learning & Clustering

Lecture 31 - CMSE 381

Prof. Elizabeth Munch

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

Mon, Dec 5, 2022

Announcements

Last time:

- Convolutional Neural Nets

This lecture:

- Clustering (Just hierarchical clustering)

20	F	Nov 4	Polynomial & Step Functions.	7.1,7.2	
21	M	Nov 7	Step Functions	7.2	
22	W	Nov 9	Basis functions, Regression Splines	7.3,7.4	
23	F	Nov 11	Decision Trees	8.1	HW #7 Due
24	M	Nov 14	Random Forests	8.2.1, 8.2.2	
25	W	Nov 16	Maximal Margin Classifier	9.1	
26	F	Nov 18	SVC	9.2	HW #8 Due
27	M	Nov 21	SVM	9.3, 9.4, 9.5	
28	W	Nov 23	Extended virtual office hours		
	F	Nov 25	No class - Thanksgiving		
29	M	Nov 28	Single layer NN	10.1	HW #9 Due
30	W	Nov 30	Multi Layer NN	10.2	
31	F	Dec 2	CNN	10.3	
32	M	Dec 5	Unsupervised Learning & Clustering	12.1, 12.4	HW #10 Due
	W	Dec 7	Review		
	F	Dec 9	Midterm #3	Bring your cheat sheet and a non-internet-connected calculator	

Announcements:

- HW #10 Due today
- Weds: Review - Bring questions!
- Friday: Exam
  - Content since 2nd Exam (Ch 7 and on)
  - One page (8.5x11) handwritten cheat sheet

Section 1

Unsupervised learning

Supervised vs Unsupervised Learning

Supervised

Unupervised

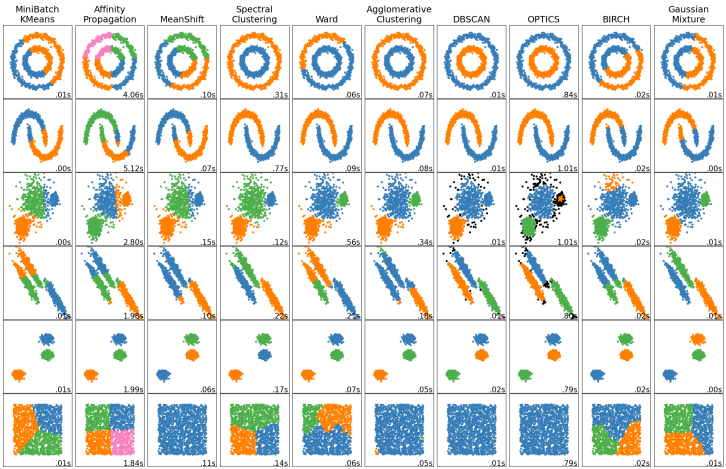
Some examples of unsupervised problems

- 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

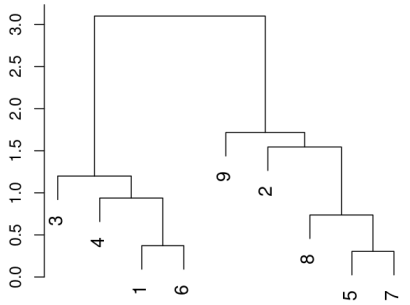
Big idea

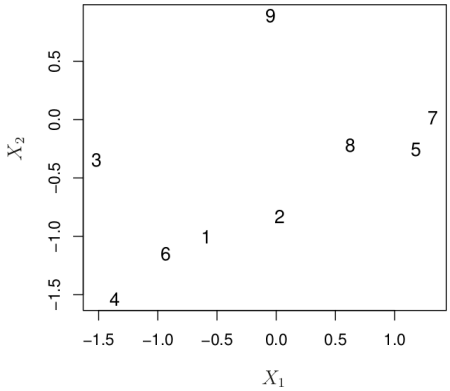

MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
OPTICS
BIRCH
Gaussian Mixture
.01s
4.06s
.10s
.31s
.06s
.07s
.01s
.84s
.02s
.01s
.00s
5.12s
.07s
.77s
.09s
.08s
.01s
1.01s
.02s
.00s
.00s
2.80s
.15s
.12s
.56s
.34s
.01s
1.01s
.02s
.01s
.01s
1.98s
.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

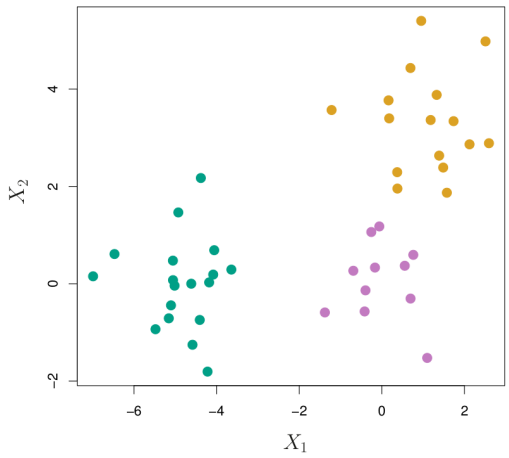
Hierarchical Clustering

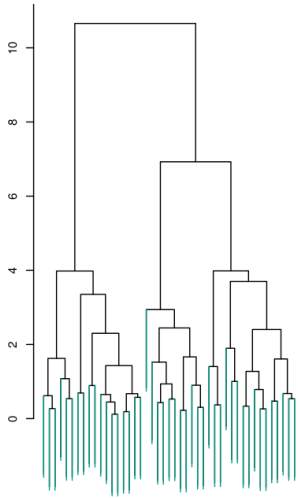
Dendrogram





A bigger example

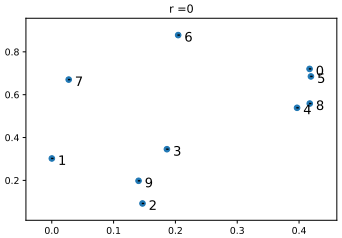


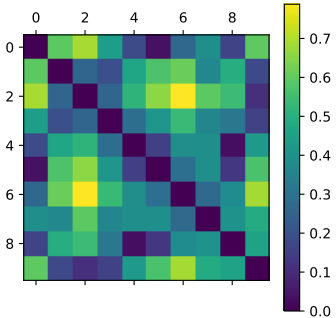


Single linkage

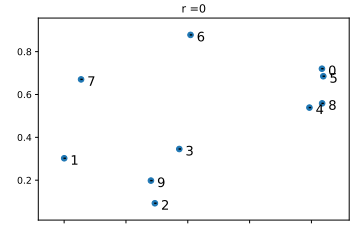
Distance between cluster A and cluster B:
Smallest distance between the points

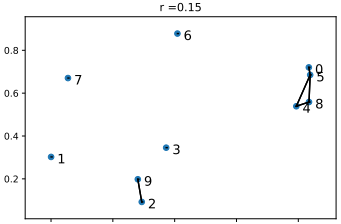
$$L(A, B) = \min_{a \in A, b \in B} \|a - b\|$$

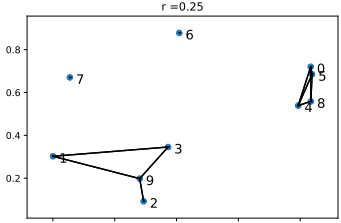


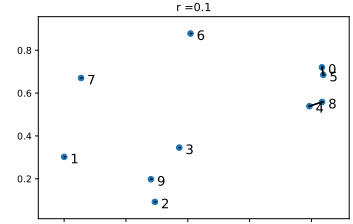


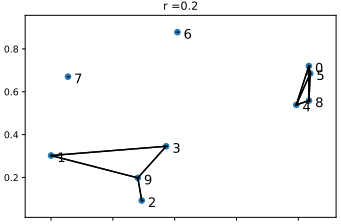
Building the dendrogram

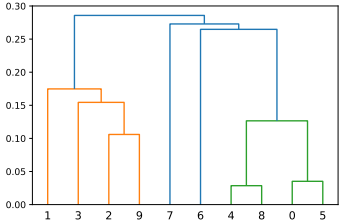




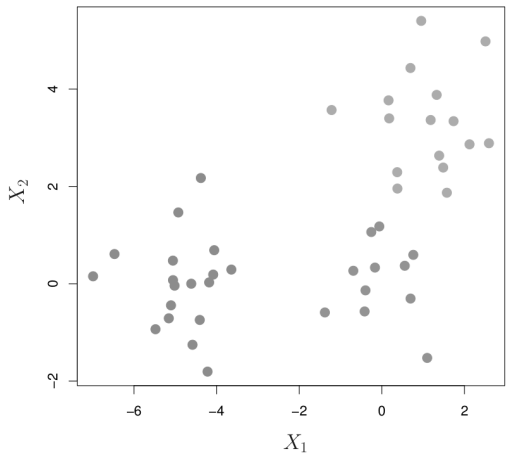


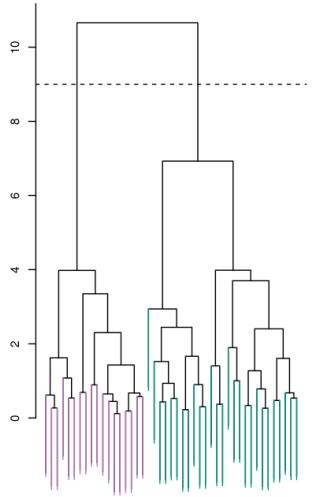




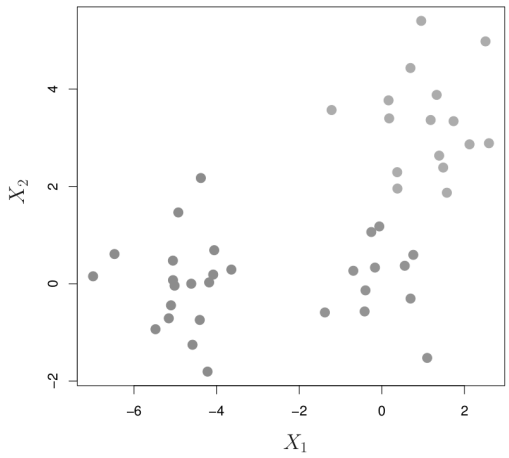


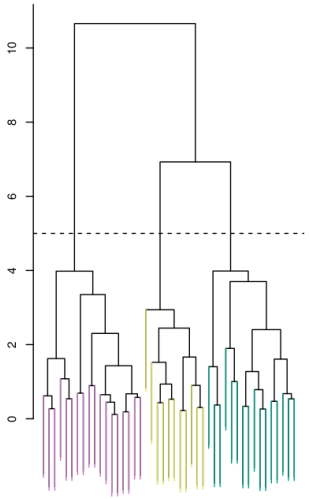
How to get clusters



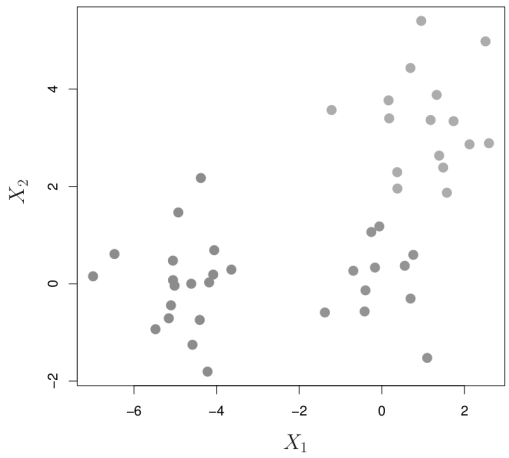


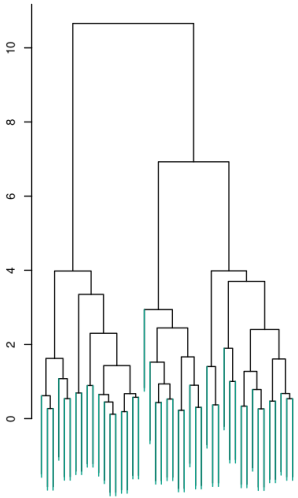
How to get different clusters





Can get any number of clusters

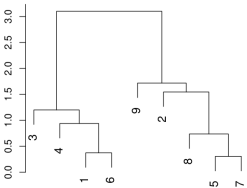




Linkage

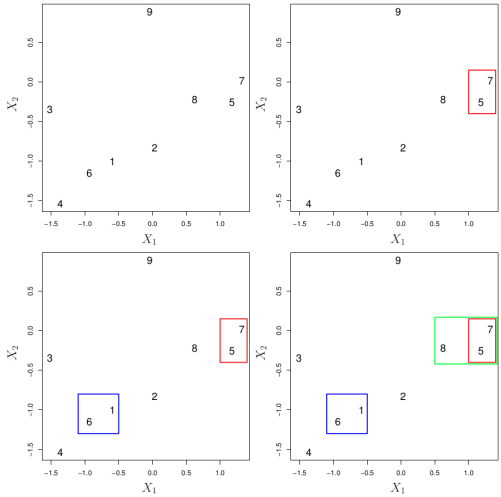
Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities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

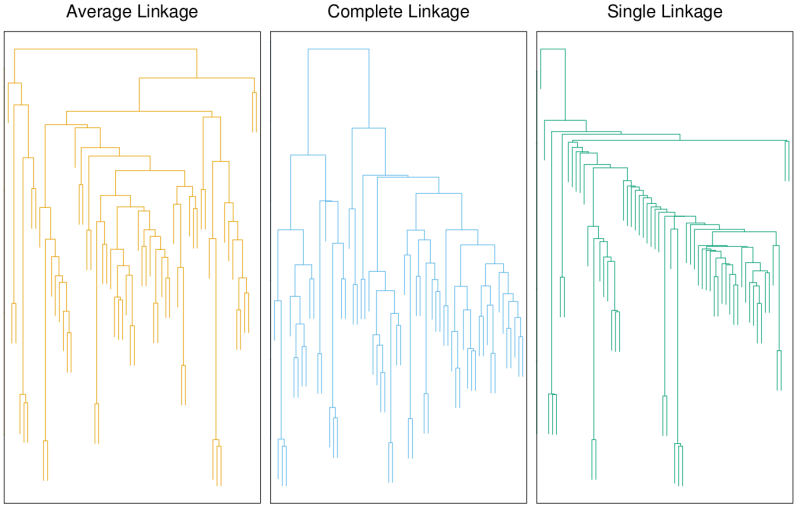


Distance between cluster A and cluster B:
Largest distance between the points

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



Examples of different linkage



Dependence on dissimilarity measure

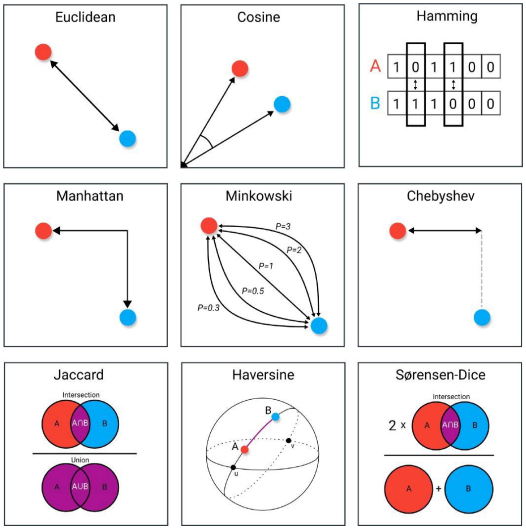


Photo Credit Link

Coding

Next time

20	F	Nov 4	Polynomial & Step Functions.	7.1,7.2	
21	M	Nov 7	Step Functions	7.2	
22	W	Nov 9	Basis functions, Regression Splines	7.3,7.4	
23	F	Nov 11	Decision Trees	8.1	HW #7 Due
24	M	Nov 14	Random Forests	8.2.1, 8.2.2	
25	W	Nov 16	Maximal Margin Classifier	9.1	
26	F	Nov 18	SVC	9.2	HW #8 Due
27	M	Nov 21	SVM	9.3, 9.4, 9.5	
28	W	Nov 23	Extended virtual office hours		
	F	Nov 25	No class - Thanksgiving		
29	M	Nov 28	Single layer NN	10.1	HW #9 Due
30	W	Nov 30	Multi Layer NN	10.2	
31	F	Dec 2	CNN	10.3	
32	M	Dec 5	Unsupervised Learning & Clustering	12.1, 12.4	HW #10 Due
	W	Dec 7	Review		
	F	Dec 9	Midterm #3	Bring your cheat sheet and a non-internet-connected calculator	