FYS: AI in Healthcare

Unsupervised Learning

John Lalor

October 2, 2018

Admin

- Assignment 3 followup

Admin

- Assignment 3 followup
- Assignment 5
    - Junior Year Writing crossover
    - Due 10/14/18 at 11:59pm (*Sunday*)

Unsupervised learning

Unsupervised learning

Used when your data is not labeled

Unsupervised learning

Used when your data is not labeled

Two types

Unsupervised learning

Used when your data is not labeled

Two types

- Clustering
    - group data together by some similarity metric

Unsupervised learning

Used when your data is not labeled

Two types

- Clustering
  - group data together by some similarity metric
- Dimensionality reduction
  - high dimension data → low dimension data

Today

Clustering

- k-means
- Hierarchical clustering
- Topic modeling

k-means clustering

Algorithm

- Initialize the k means
- Until convergence:
    - Assign data to clusters based on the closest mean (Euclidean distance)
    - Recalculate means using cluster assignments

k-means demo

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

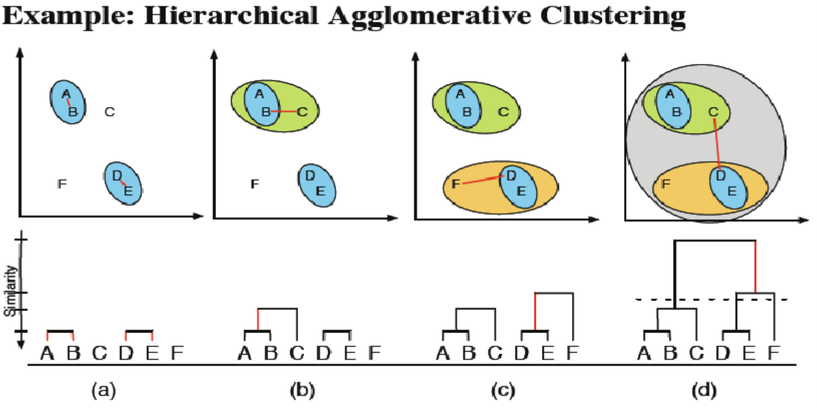
Hierarchical clustering

Two types: bottom-up (“agglomerative”) and top-down (“divisive”)

Janssen, Peter et al. Cluster analysis to understand socio-ecological systems: a guideline.

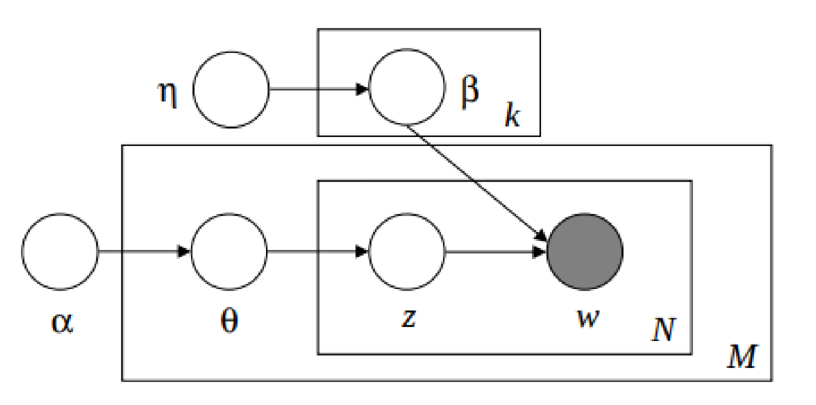
Hierarchical clustering

Two types: bottom-up ("agglomerative") and top-down ("divisive")



Janssen, Peter et al. Cluster analysis to understand socio-ecological systems: a guideline.

Topic modeling



Topic modeling

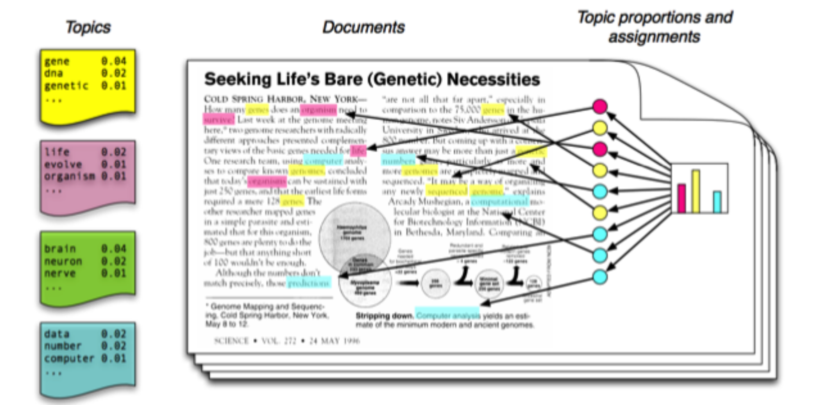


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Next week

Dimensionality reduction

Activity: Nearest Neighbors

Warm-up

Training set: 1, 7, 13

Test set: 3, 17

$k = 1$

Round 1

Training set: 1, 2

Test set: 14, 16

$k = 1$

Round 1

Training set: 1, 2

Test set: 14, 16

$k = 1$

Training data is important

Round 2

Training set: 5, 15

Test set: 4, 12

$k = 2$

Round 2

Training set: 5, 15

Test set: 4, 12

$k = 2$

Use an odd $\#$ of neighbors

Round 3

Training set: 3

Test set: 13

$k = 1$

Round 3

Training set: 3

Test set: 13

$k = 1$

“Nearest” might not be very close

Round 4

Training set: 17

Test set: 7

$k = 1$

Round 4

Training set: 17

Test set: 7

$k = 1$

“Nearest” might not be very close

Round 5

Training set: 8, 10, 11, 12, 18

Test set: 9, 19

$k = 3$

Round 5

Training set: 8, 10, 11, 12, 18

Test set: 9, 19

$k = 3$

KNN is slow at test time

Round 6

Training set: 6

Test set: 7

$k = 1$

Round 6

Training set: 6

Test set: 7

$k = 1$

Be mindful of outliers!