Unsupervised Machine Learning

- PCA
- Clustering
- Sequential mining

1

Unsupervised Learning

- Dimensionality Reduction
  - PCA
  - t-SNE (t-distributed Stochastic Neighbor Embedding)
  - Factor analysis
- Clustering
  - k-means
  - Hierarchical
  - Soft clustering (Gaussian mixture)

2

Dimensionality Reduction

- Complexity of model ~ dimensionality (# of features)
  - Dimensionality vs. sample size
  - Simpler models have lower variance
  - Simpler models are more explainable

- Reducing dimensionality
  - Feature selection
  - Feature extraction

3

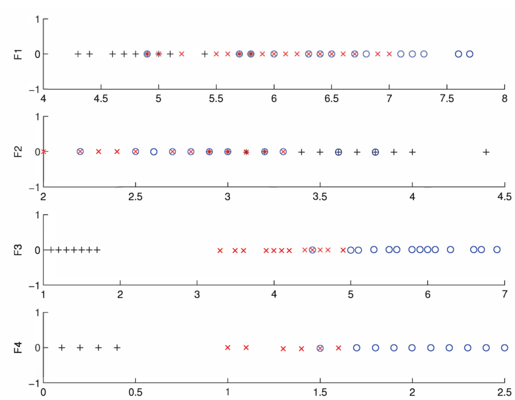
Feature Selection

- Feature selection approaches
  - Forward selection
  - Backward selection

- Given d features
  - There are 2^d collections of features

    $d = 10 \rightarrow$ 1,000 choices

    $d = 20 \rightarrow$ 1,000,000 choices

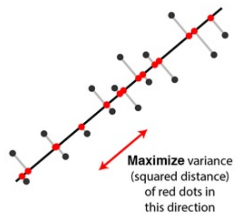


Iris data set

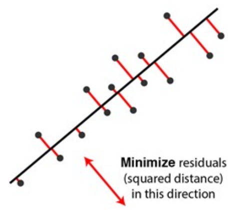
4

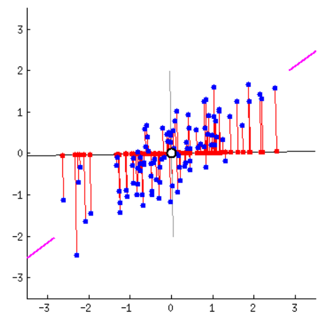
Feature Extraction

- PCA (Principal Component Analysis)
  - Dimensionality reduction technique
  - Significance of data should remain
    - Maximize the variance in the reduced space
    - Unlike, minimization of projection error



vs





UNIVERSITY OF TORONTO | Engineering

5

Principal Component Analysis

- Projection of x on the direction of w

$$z = w^T x$$

- Maximizing the variability subject to size of vector = 1

$$\max_{w_1} \; Var(z_1) = w_1^T \Sigma w_1$$
$$\text{s.t.} \quad w_1^T w_1 = 1$$

- By Lagrangian

$$\max_{w_1} \; w_1^T \Sigma w_1 - \alpha(w_1^T w_1 - 1)$$

- By first order condition

$$2\Sigma w_1 - 2\alpha w_1 = 0 \qquad \text{or} \qquad \Sigma w_1 = \alpha w_1$$

UNIVERSITY OF TORONTO | Engineering

6

Principal Component Analysis

- We have

$$\Sigma w_1 = \alpha w_1$$

  - w_1 is an eigenvector of Σ and α is the corresponding eigenvalue

- Since we try to maximize $w_1^T \Sigma w_1 = \alpha w_1^T w_1 = \alpha$, we choose the eigenvector with the **largest eigenvalue**

- **The second PCA** w_2 should also maximize the variance, be of unit length and be orthogonal to w_1

$$\max_{w_2} \; w_2^T \Sigma w_2 - \alpha(w_2^T w_2 - 1) - \beta(w_2^T w_1 - 0)$$

- By first order condition

$$2\Sigma w_2 - 2\alpha w_2 - \beta w_1 = 0$$

- Multiplying w_1

$$2w_1^t \Sigma w_2 - 2w_1^t \alpha w_2 - \beta w_1^t w_1 = 0$$

UNIVERSITY OF TORONTO Engineering

7

Principal Component Analysis

- From the previous slide,

$$2w_1^T \Sigma w_2 - 2w_1^T \alpha w_2 - \beta w_1^T w_1 = 0 \qquad (1)$$

- Since

$$w_1^T w_2 = 0 \qquad \Sigma w_1 = \lambda_1 w_1$$

- $w_1^T \Sigma w_2$ is a scalar and hence its transpose has the same the value (i.e., $w_1^T \Sigma w_2 = w_2^T \Sigma w_1$). Therefore,

$$w_1^T \Sigma w_2 = w_2^T \Sigma w_1 = \lambda_1 w_2^T w_1 = 0$$

- (1) becomes $0 - 0 - \beta \cdot 1 = 0$ implying $\beta = 0$. As a result,

$$2\Sigma w_2 - 2\alpha w_2 - \beta w_1 = 0 \quad \text{becomes} \quad \Sigma w_2 = \alpha w_2$$

- Similar to the first PCA, the second PCA is the eigenvector of Σ with the **second largest eigenvalue**

UNIVERSITY OF TORONTO Engineering

8

Steps of PCA

1. Standardization/normalization of Data
2. Computation of Covariance Matrix Σ
3. Computation of eigenvector and eigenvalues
4. Selection of Principal Components out of eigenvectors
5. Projection of the data using the principal components vectors

9

- Data and Standardization

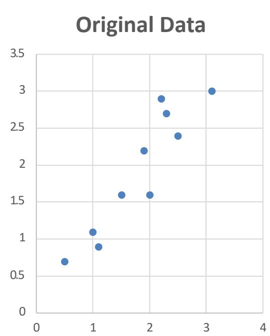
Data =

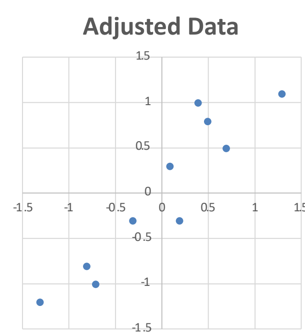
x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

$$(x', y') = (x - x, y - y)$$

DataAdjust =

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01





10

- Covariance matrix Σ

$$\Sigma = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

x	y
.69	.49
-1.31	-1.21
.39	.99
.09	.29
1.29	1.09
.49	.79
.19	-.31
-.81	-.81
-.31	-.31
-.71	-1.01

	(X'-Avg(X))	Y'-Avg(Y)	(X'-Avg(X))(Y'-Avg(Y))
	0.69	0.49	0.3381
	-1.31	-1.21	1.5851
	0.39	0.99	0.3861
	0.09	0.29	0.0261
	1.29	1.09	1.4061
	0.49	0.79	0.3871
	0.19	-0.31	-0.0589
	-0.81	-0.81	0.6561
	-0.31	-0.31	0.0961
	-0.71	-1.01	0.7171
Sum			
Sum of squares	5.549	6.449	5.539
Sample Variance	**0.61655556**	**0.71655556**	**0.615444444**

UNIVERSITY OF TORONTO | Engineering

11

- Eigenvalues

Given $\Sigma = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$

We want to find λ such that $\det(\Sigma - \lambda I) = 0$

UNIVERSITY OF TORONTO | Engineering

12

- Given eigenvalues, let's find eigenvectors $\boldsymbol{v}$

$$\Sigma v = \lambda v$$
$$(\Sigma - \lambda I)v = 0$$

For $\lambda = 1.28402771$

$$\begin{bmatrix} -0.677873399 \\ -0.735178656 \end{bmatrix}$$

For $\lambda = 0.0490834$

$$\begin{bmatrix} -0.735178656 \\ -0.677873399 \end{bmatrix}$$

UNIVERSITY OF TORONTO | Engineering

13

- Eigenvectors

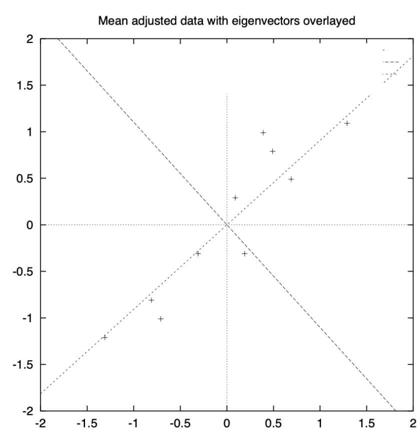
$$\begin{bmatrix} -0.735178656, -\mathbf{0.677873399} \\ -0.677873399, -\mathbf{0.735178656} \end{bmatrix}$$

- PCAs (feature vectors)
  - First component

$$\begin{bmatrix} -0.677873399 \\ -0.735178656 \end{bmatrix}$$

  - Second component

$$\begin{bmatrix} -0.735178656 \\ 0.677873399 \end{bmatrix}$$



UNIVERSITY OF TORONTO | Engineering

14

- Projected Data

	x	y
DataAdjust =	.69	.49
	-1.31	-1.21
	.39	.99
	.09	.29
	1.29	1.09
	.49	.79
	.19	-.31
	-.81	-.81
	-.31	-.31
	-.71	-1.01

$(\tilde{x}, \tilde{y}) = Feacture\ Vector \cdot (x, y)$

$\tilde{x}$
-.827970186
1.77758033
-.992197494
-.274210416
-1.67580142
-.912949103
.0991094375
1.14457216
.438046137
1.22382056

$$\begin{pmatrix} -.677873399 & -.735178656 \\ -.735178656 & .677873399 \end{pmatrix}$$

$\tilde{x}$	$\tilde{y}$
-.827970186	-.175115307
1.77758033	.142857227
-.992197494	.384374989
-.274210416	.130417207
-1.67580142	-.209498461
-.912949103	.175282444
.0991094375	-.349824698
1.14457216	.0464172582
.438046137	.0177646297
1.22382056	-.162675287

UNIVERSITY OF TORONTO Engineering

15

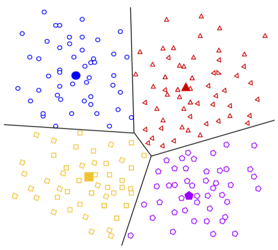
Clustering

- A process of grouping a set of data objects into groups or clusters
  - So that objects within a cluster have high similarity,
  - But are dissimilar to objects in other clusters.
- An unsupervised ML
- Applications
  - Anomaly detection
  - Marketing (segmentation)
  - Recommender system
  - Social network analysis

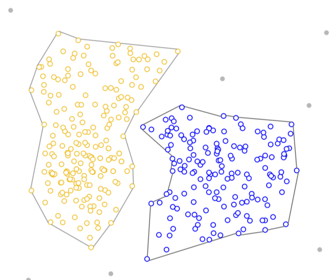
UNIVERSITY OF TORONTO Engineering

17

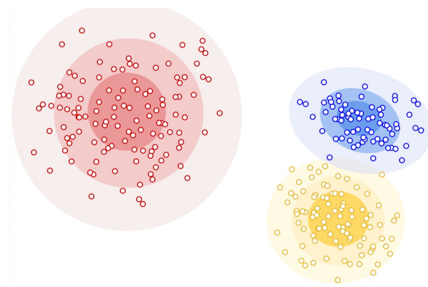
Clustering



Centroid-based



Density-based

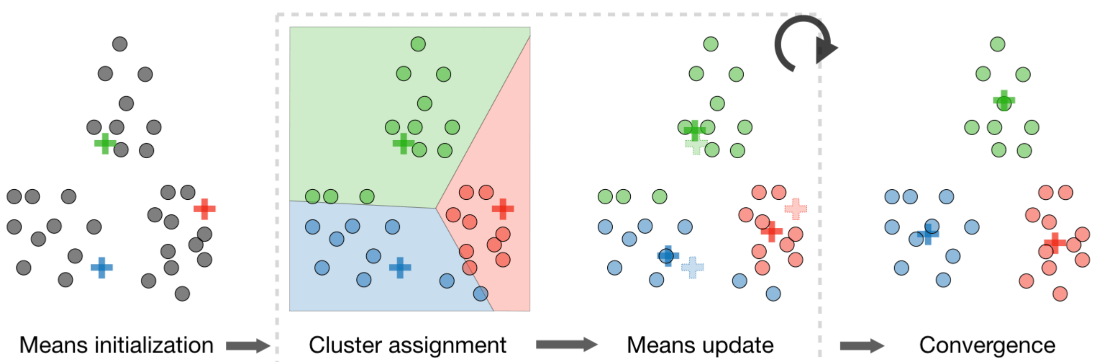


Distribution-based

18

Hard Clustering

- k-Means Clustering



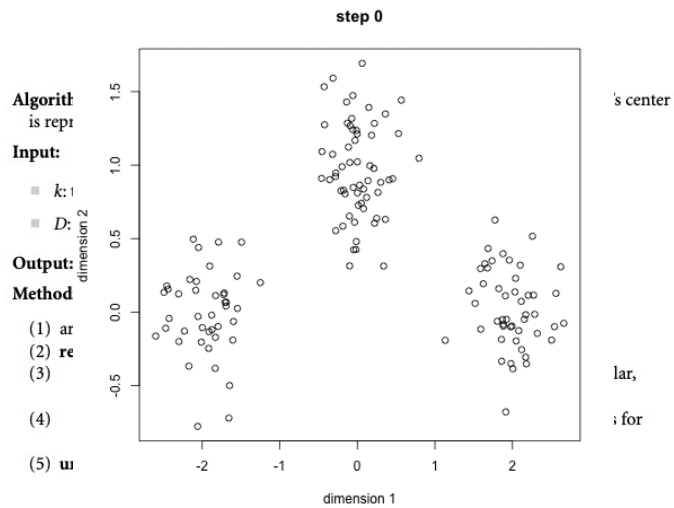
19

k-Means Clustering

- Error

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - c_i)^2$$

- Drawbacks
  - # of clusters needs to be given
  - Final results depend on the initial random selection of cluster centres
  - Sensitive to outliers (k-Medoids)



UNIVERSITY OF TORONTO | Engineering

20

Clustering

- Determining the # of Clusters
  - To balance between compressibility & accuracy
  - Depends on distribution's shape, scale of the data set, required resolution
  - Simple Rule

$$k = \sqrt{\frac{n}{2}}$$

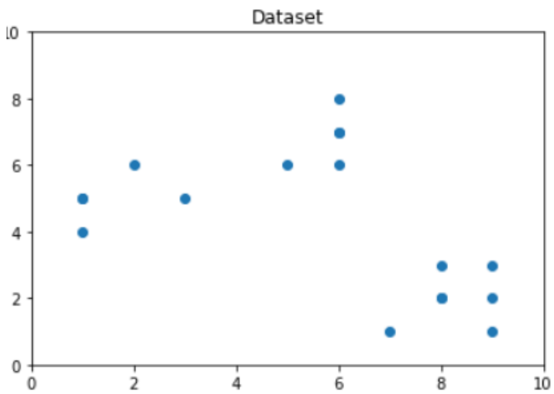
  - In which case each cluster has $\sqrt{2n}$ points
  - The Elbow Method
    - More clusters means smaller within-cluster variance
    - Marginal variance reduction decreases
    - Hence, we can heuristically look for a turning point

```
1: Form k clusters
2: Compute the sum of within-
   cluster error (k)
3: Plot the curve of error w.r.t. k
4: Find the first turning point in
   the curve
```

UNIVERSITY OF TORONTO | Engineering

21

The Elbow Method

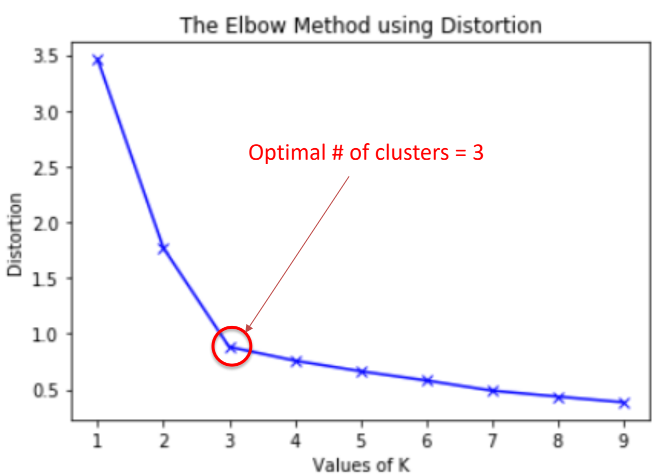


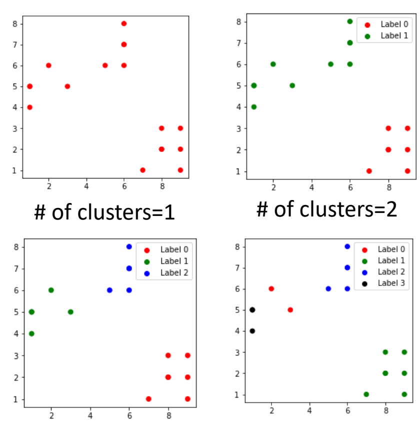
k	Sum of inner cluster variance
1	: 3.4577032384495707
2	: 1.7687413573405673
3	: 0.8819889697423957
4	: 0.7587138847606585
5	: 0.6635212812400347
6	: 0.5808803063754726
7	: 0.5093717077076824
8	: 0.41652236641410356
9	: 0.3333333333333333

UNIVERSITY OF TORONTO | Engineering

22

The Elbow Method





# of clusters=1   # of clusters=2

# of clusters=3   # of clusters=4

UNIVERSITY OF TORONTO | Engineering

23

Case Study

Data

- 480 generating units
  - Hydro power plant in Niagara Falls, Canada
  - Over 5 years (2012~2017)
- The units failed for various causes from 114 components.
- 0.6 million entries of maintenance records, failures, etc.

472-megawatt steam turbine generator (photo credit: businesswire.com)

UNIVERSITY OF TORONTO | Engineering

24

Data Collection

1. Removing redundancy
2. Removing units with inadequate records
3. Removing or recovering incomplete/inconsistent observations

	ForceOut	MainOut	MaxCapability	NumofCommon	PlanOut	WorkingHour	G199999	G142100	G141100	G142115	...
HGU0001	11	13	77.0	0.0	12	33429.6	0.0	6.0	2.0	0.0	...
HGU0002	13	14	77.0	0.0	10	36574.6	0.0	9.0	1.0	0.0	...
HGU0003	5	14	77.0	0.0	6	35721.3	0.0	3.0	1.0	0.0	...
HGU0004	11	10	77.0	0.0	7	36983.2	0.0	4.0	0.0	1.0	...
HGU0005	11	13	77.0	0.0	9	39080	0.0	6.0	1.0	0.0	...
HGU0006	6	13	77.0	0.0	8	35225.6	0.0	6.0	2.0	0.0	...
HGU0007	5	4	150.0	0.0	7	40213.8	1.0	7.0	16.0	1.0	...

UNIVERSITY OF TORONTO | Engineering

25

Clustering Results

	Cluster 1	Cluster 2	Cluster 0
Average number of Forced outages	25.985	13.603	17.460
Average number of Maintenance outages	30.758	15.026	23.400
Average number of Planned outages	15.833	10.250	26.100
Average number of Common modes	0.015	1.263	0.280
Average maximum capability	46.533	58.185	306.586
Average working hours	37738.700	38917.044	35084.975

UNIVERSITY OF TORONTO | Engineering

26

Clustering Results

Cluster 0

- Largest average maximum capacity/unit
- Medium reliability
- Highest planned outages number is scheduled on these units
- Cluster 0 seems mostly important to the company.

Cluster 1

- Smallest average maximum capability/unit
- The least reliable units
- Relatively large number of planned outages
- Failures seem isolated rather than caused by other units

Cluster 2

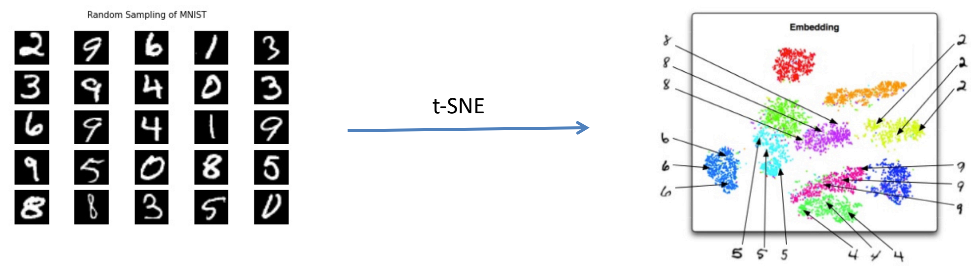
- The most reliable unit
- Outages are likely caused by other units (leading to correlation analysis)
- Least maintained

UNIVERSITY OF TORONTO | Engineering

27

t-SNE

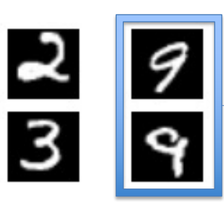
- t-Distributed Stochastic Neighbor Embedding



  - Stochastic tool to retain local relationship through dimensionality reduction
  - Distance between clusters is not controlled
  - Computationally expensive

t-SNE

- PCA: Push points away from each other
- t-SNE: Keep closer points closer



- Nearness measured by probability that x_i belongs to the neighborhood of x_j than others

$$p_{i|j} := \frac{\exp(-|x_i - x_j|^2/2\sigma_j^2)}{\sum_{k \neq j} \exp(-|x_k - x_j|^2/2\sigma_j^2)}$$

- Keep the probability high in the reduced space

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

t-SNE

- Steps
  1. Construct a probability distribution on pairs in higher dimensions

$$p_{ij} = \frac{p_{ji} + p_{ij}}{2N}$$

  2. Construct a probability distribution on pairs in the reduced dimensions

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

  3. Use SGD to minimize the discrepancy between the two distributions

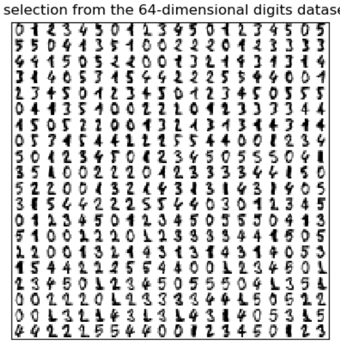
$$\mathrm{KL}\,(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

UNIVERSITY OF TORONTO | Engineering

30

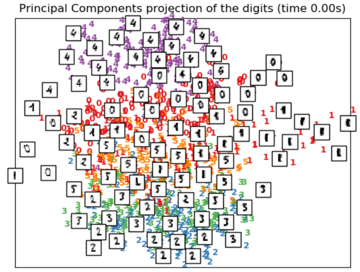
t-SNE vs. PCA

A selection from the 64-dimensional digits dataset



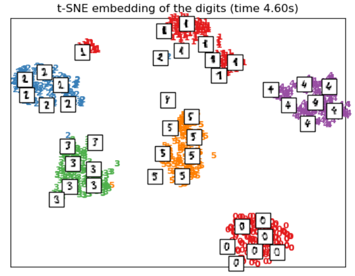
a. PCA

Principal Components projection of the digits (time 0.00s)



b. t-SNE

t-SNE embedding of the digits (time 4.60s)

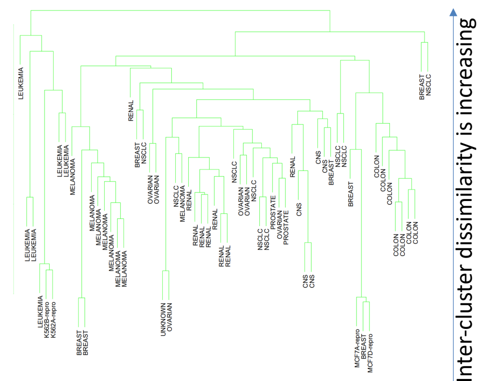


UNIVERSITY OF TORONTO | Engineering

31

Hierarchical Clustering

- Avoid the critical hyper-parameter: k
  - Instead, *threshold* on (inter-group) dissimilarity
- Two approaches
  - Bottom up (agglomerative)
  - Top-down (divisive)
- Both possess "monotonicity property"
- Disjoint clusters are defined by cutting the dendrogram horizontally
- Dendrogram depends on dissimilarity metric



dendrogram

33

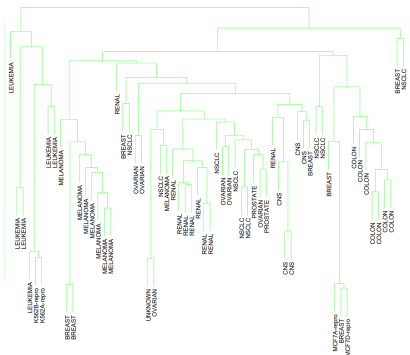
Hierarchical Clustering

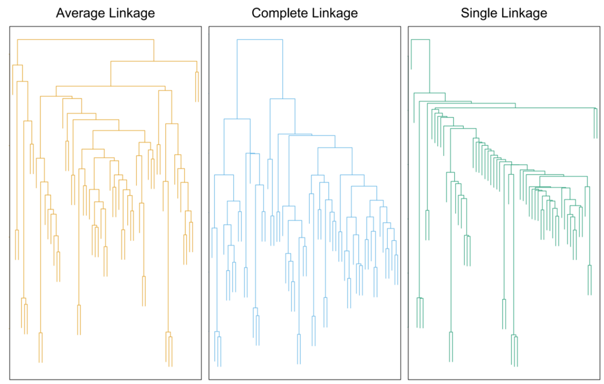
- Agglomerative Clustering
  - Begins with clusters of a single sample
  - The closest two clusters are merged; and move upward to continue
- Dissimilarity metric
  - Given two groups G and H, the dissimilarity between G and H is $d(G, H)$
  - Dissimilarity between sample $i \in G$ and sample $i' \in H$ is $d_{ii'}$
  - Single linkage-based $$d_{SL}(G, H) = \min_{\substack{i \in G \\ i' \in H}} d_{ii'}$$
  - Complete linkage-based $$d_{CL}(G, H) = \max_{\substack{i \in G \\ i' \in H}} d_{ii'}$$
  - Average linkage-based $$d_{GA}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'}$$

34

Hierarchical Clustering

- 3 dendrograms using three metrics





UNIVERSITY OF TORONTO | Engineering

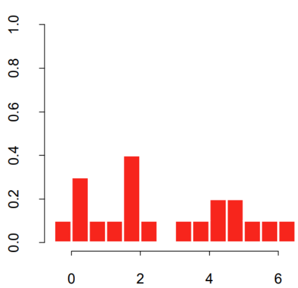
35

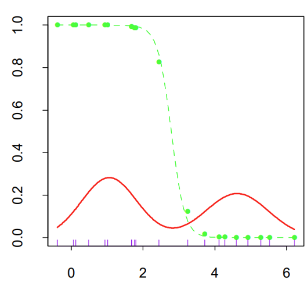
Soft-Clustering

- Histogram and 2-component Gaussian mixture
  - 20 samples

-0.39	0.12	0.94	1.67	1.76	2.44	3.72	4.28	4.92	5.53
0.06	0.48	1.01	1.68	1.80	3.25	4.12	4.60	5.28	6.22

  - Visualization





UNIVERSITY OF TORONTO | Engineering

36

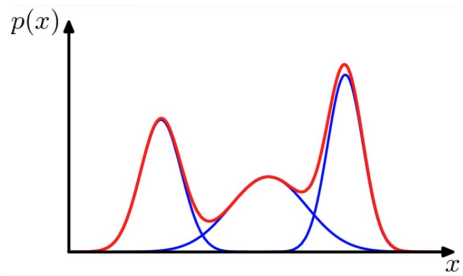
Soft-Clustering

- Clustering ≡ Mixture Distribution Fitting
- Data distribution can be

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

cluster membership (responsibilities)

data distribution in cluster k



- Log likelihood can be maximized
  - Usually, it is intractable
  - Alternating (1) find membership and (2) fit a distribution in each cluster: EM Algorithm

UNIVERSITY OF TORONTO Engineering

37

Sketch of Derivation

- Responsibilities as a probability distribution
  - Let $z \in \{0,1\}^K$ such that $z_k \in \{0,1\}$ and $\sum_k z_k = 1$
  - Let $p(z_k = 1) = \pi_k$ $(0 \le \pi_k \le 1,\ \sum_{k=1}^{K} \pi_k = 1)$
  - We can write this distribution as

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- Conditional distribution of $\boldsymbol{x}$ given a particular value of $\boldsymbol{z}$

$$p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k)$$

Hence,

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})^{z_k}$$

UNIVERSITY OF TORONTO Engineering

38

- Posterior distribution of responsibilities z_k

$$\gamma(z_k) \equiv p(z_k = 1|x) = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)}$$
$$= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}$$

UNIVERSITY OF TORONTO | Engineering

39

EM for Gaussian Mixtures

1. Initialize the means μ_k, covariances Σ_k and mixing coefficient π_k
2. (**Expectation**) Compute the responsibilities

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

3. (**Maximization**) Update the parameters given the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \pi_k^{\text{new}} = \frac{N_k}{N}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \qquad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

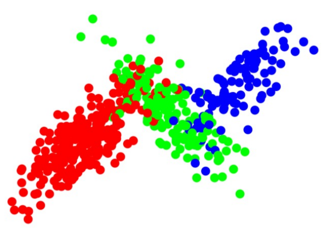
4. If converged, stop; otherwise, go to Step 2

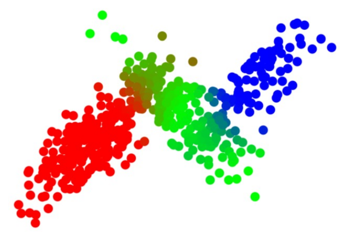
UNIVERSITY OF TORONTO | Engineering

40

Soft-clustering

Steps	k-means	Soft-clustering
Parameter computation	Compute new **centroids** given hard allocation	Compute **mean & variance** of component Gaussian distributions given soft allocation
Cluster allocation	Assign samples to the nearest centroid (hard allocation)	Assign samples to all clusters with fractional membership (soft allocation)





UNIVERSITY OF TORONTO | Engineering

41