NYU

Introduction to Data Science

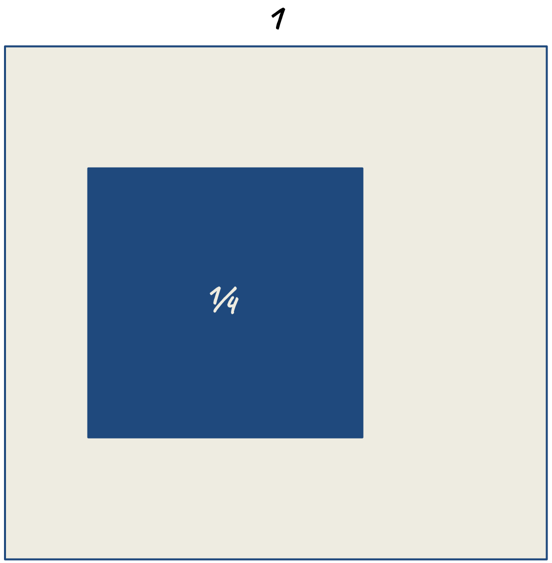
Center for Data Science
Iddo Drori, Spring 2019

Agenda

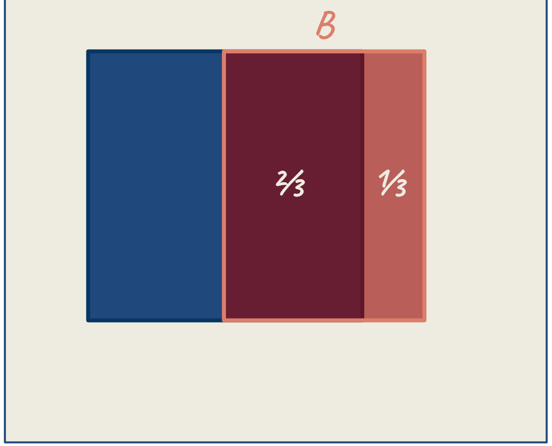
- Bayes rule, K-nearest neighbors for classification and regression
- Unsupervised learning, clustering
  - K-means clustering
  - Hierarchical clustering
  - Clustering quality using mutual information
- Dimensionality reduction
  - Principal component analysis (PCA)
  - t-SNE

Bayes Rule

Conditional Probability

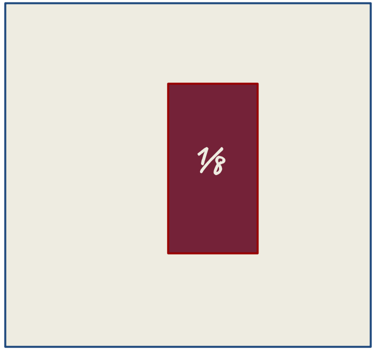


$P(A) = 1/4$

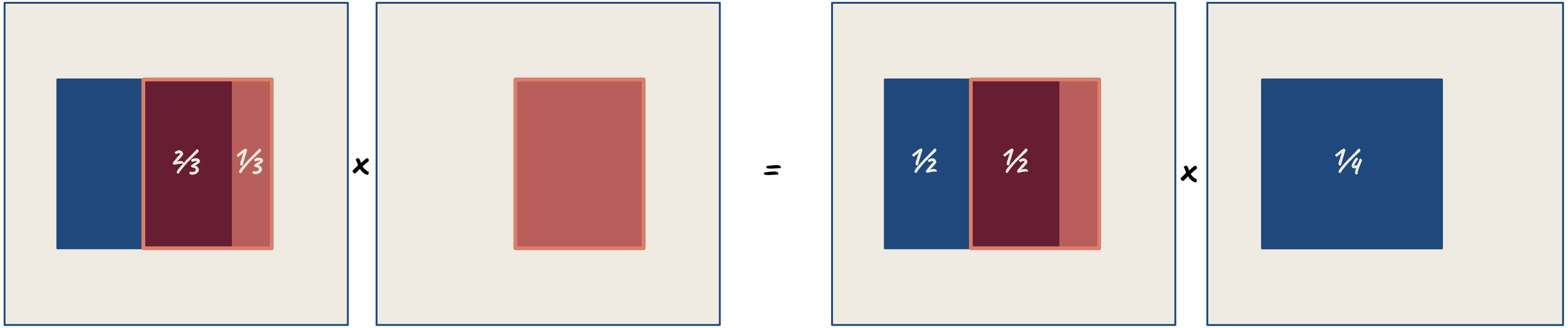


$P(A|B) = 2/3$

Product Rule



$P(A|B)P(B) \quad = \quad P(A,B) \quad = \quad P(B|A)P(A)$



$P(A|B)=2/3$ $\quad$ $P(B)=3/16$ $\quad$ $P(B|A)=1/2$ $\quad$ $P(A)=1/4$

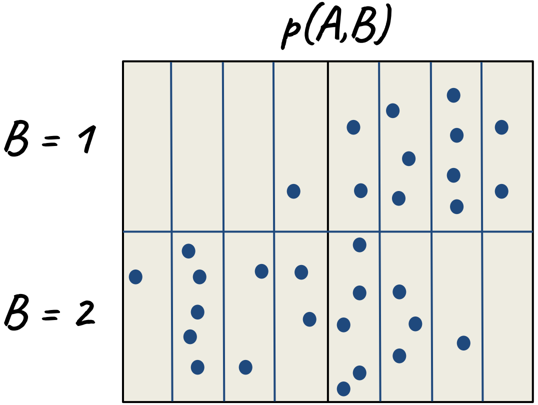
Bayes Rule

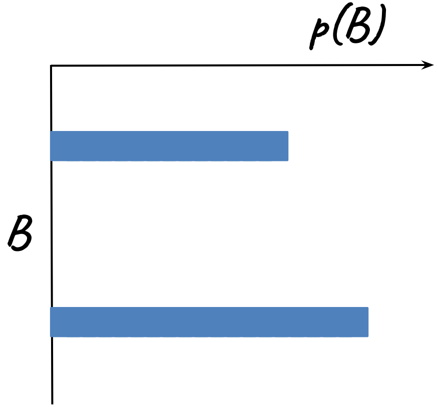
$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(B|A) = P(A|B)P(B) \,/\, P(A)$$

Sum Rule

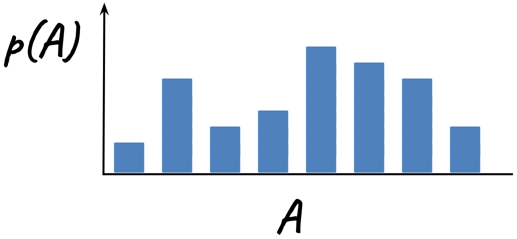
$$p(A) = \sum_{B} p(A,B)$$

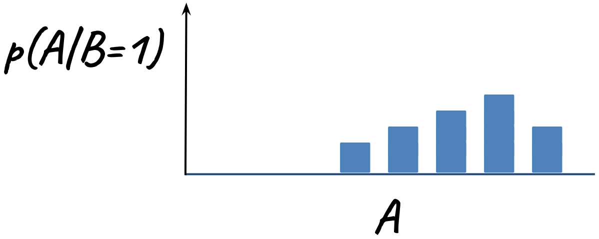




$$\sum_{i=1}^{2} p(B=i) = 1$$

$$\sum_{j=1}^{8} p(A=j) = 1$$





Bayes Rule

$$P(B_1|A) = P(A|B_1)P(B_1) \,/\, P(A)$$

$$P(B_2|A) = P(A|B_2)P(B_2) \,/\, P(A)$$

$$\frac{P(B_1|A)}{P(B_2|A)} = \frac{P(A|B_1)}{P(A|B_2)} \times \frac{P(B_1)}{P(B_2)}$$

posterior odds ratio is likelihood ratio times prior ratio

Bayes Rule

	same gender	different gender
identical		
fraternal		

P(identical) = ⅓

P(fraternal) = ⅔

twin girls

Observe twin girls and know that ⅓ of twin births are identical.

Question: What is the probability that the twin girls are identical?

Bayes Rule

	same gender	different gender	
identical	$1/3$	0	$P(\text{identical}) = 1/3$
fraternal	?		$P(\text{fraternal}) = 2/3$
	twin girls		

Bayes Rule

	same gender	different gender	
identical	1/3	0	P(identical) = 1/3
fraternal	1/3	1/3	P(fraternal) = 2/3

twin girls

P(identical|same gender) = P(same gender|identical)P(identical)/P(same gender) = 1 x 1/3 / P(same gender)
P(fraternal|same gender) = P(same gender|fraternal)P(fraternal)/P(same gender) = 1/2 x 2/3 / P(same gender)

Bayes Rule

	same gender	different gender	
identical	1/3	0	P(identical) = 1/3
fraternal	1/3	1/3	P(fraternal) = 2/3
	twin girls		

$$\frac{P(\text{identical}|\text{same gender})}{P(\text{fraternal}|\text{same gender})} = \frac{1/3}{1/2 \times 2/3} = 1$$

Bayes Rule

	same gender	different gender	
identical	1/3	0	P(identical) = 1/3
fraternal	1/3	1/3	P(fraternal) = 2/3
	twin girls		

P(identical|same gender) = P(fraternal|same gender)

Answer: probability that twin girls are identical is ½

Supervised Learning
Classification and Regression
K-Nearest Neighbors

K-Nearest Neighbors

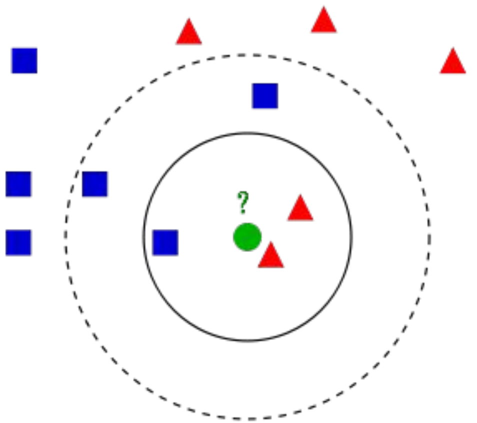
Simple and efficient algorithm

Requires definition of distance function or similarity between samples

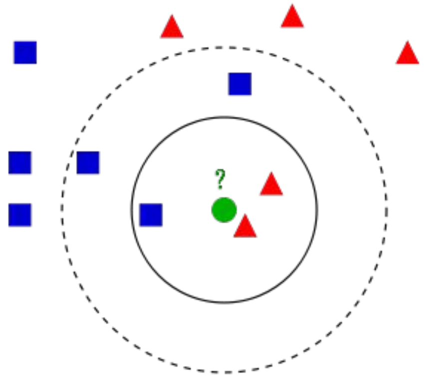
Select class based on majority vote of k closest points

Choice of k determines smoothness of classifier

Probabilistic view: approximate Bayes rule on subset



Bayes Rule: K-Nearest Neighbors



x new data point to classify

y class

V selected ball

P probability that random point is in V

N total number of samples (11)

K number of nearest neighbors (3)

N_1 total number of samples from class 1 (5)

K_1 number of samples from class 1 in V (2)

$$p(x|y=1) = \frac{K_1}{N_1}$$

$$p(y=1) = \frac{N_1}{N}$$

$$p(x) = \frac{K}{N}$$

Bayes rule: $p(y=1|x) = \dfrac{\overset{2/5}{p(x|y=1)}\,\overset{5/11}{p(y=1)}}{\underset{3/11}{p(x)}} = \overset{2/3}{\dfrac{K_1}{K}}$

K-Nearest Neighbors

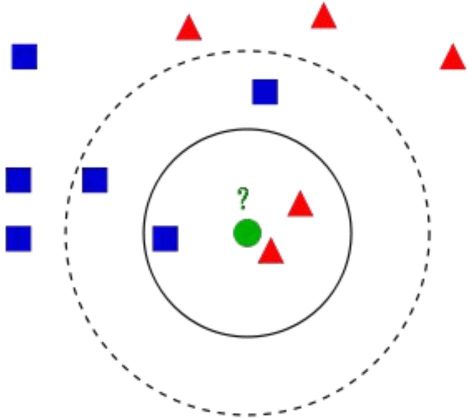
Used both for classification and regression

Explainability: show k-NN

No training, testing computation.

Larger k results in higher bias

Smaller k results in higher variance



Similarity and Distance

- Data as points in high dimensional space

- Distance between points:
  - Euclidean $\sqrt{\sum (x_i - y_i)^2}$
  - Manhattan $\sum |x_i - y_i|$
  - Jaccard $d_J(X,Y) = 1 - J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$
  - Cosine $d_C(X,Y) = 1 - C(X,Y) = 1 - \cos(\theta) = 1 - \frac{X \cdot Y}{\|X\|_2 \|Y\|_2}$

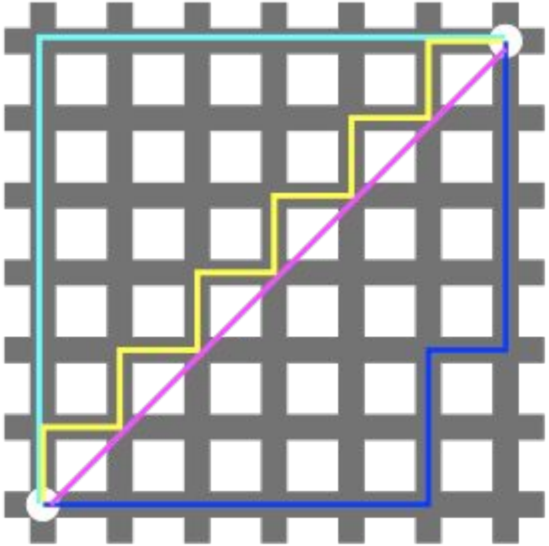
Similarity and Distance

- Feature ranges for similarity: scaling, bins, different metrics
- Irrelevant features: selection

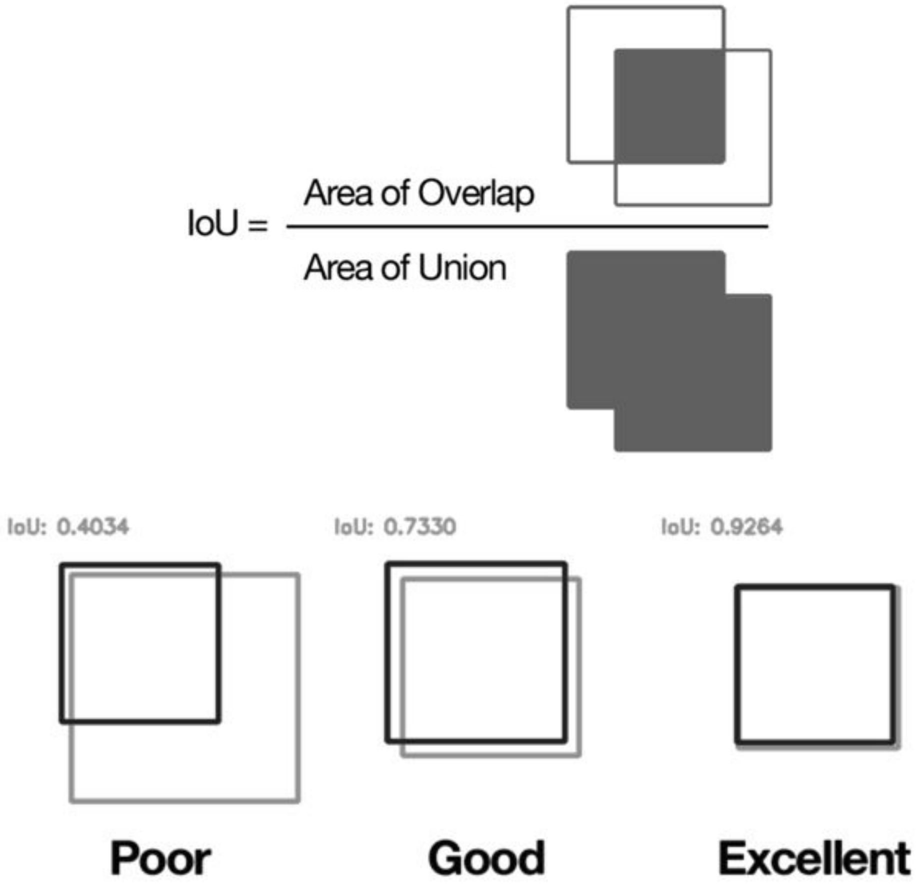
L2 and L1 Distances

- Euclidean distance $\sqrt{\sum (x_i - y_i)^2}$

- Manhattan distance $\sum |x_i - y_i| = \sum |x_i - y_i| = \sum |x_i - y_i|$



Jaccard Index



Cosine Distance

- Used in text classification
- Coefficients can be word counts

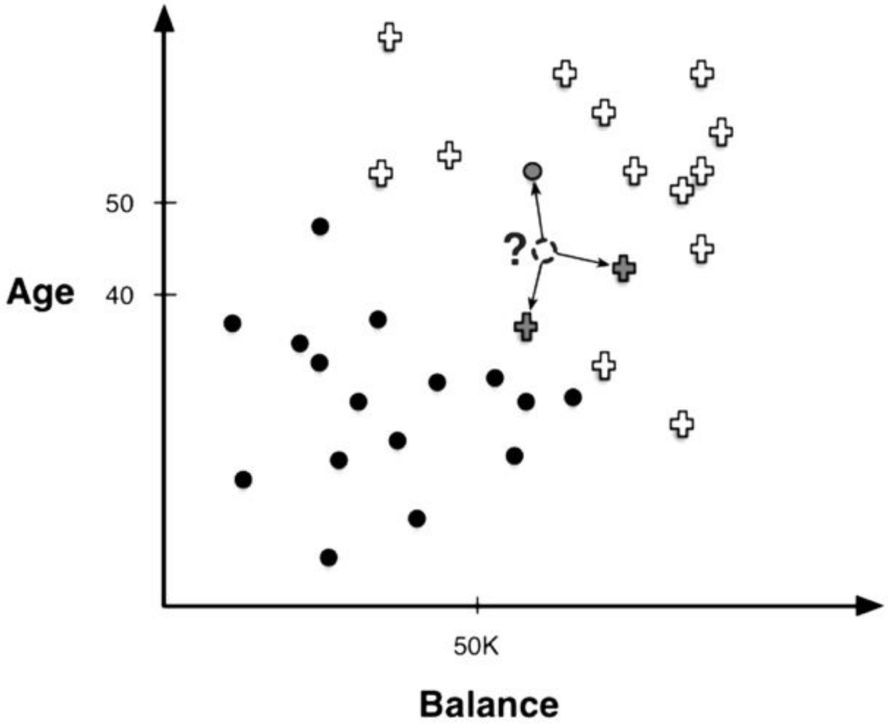
$$d_{cosine}(\mathbf{X},\ \mathbf{Y}) = 1 - \frac{\mathbf{X} \cdot \mathbf{Y}}{\| \mathbf{X} \|_2 \cdot \| \mathbf{Y} \|_2}$$

Edit Distance: Levenshtein Metric

- Count minimum number of edit operations converting between strings
- Operations: insert, delete, replace
- Sequences where order is important

K-Nearest Neighbors

- Classification
- Regression



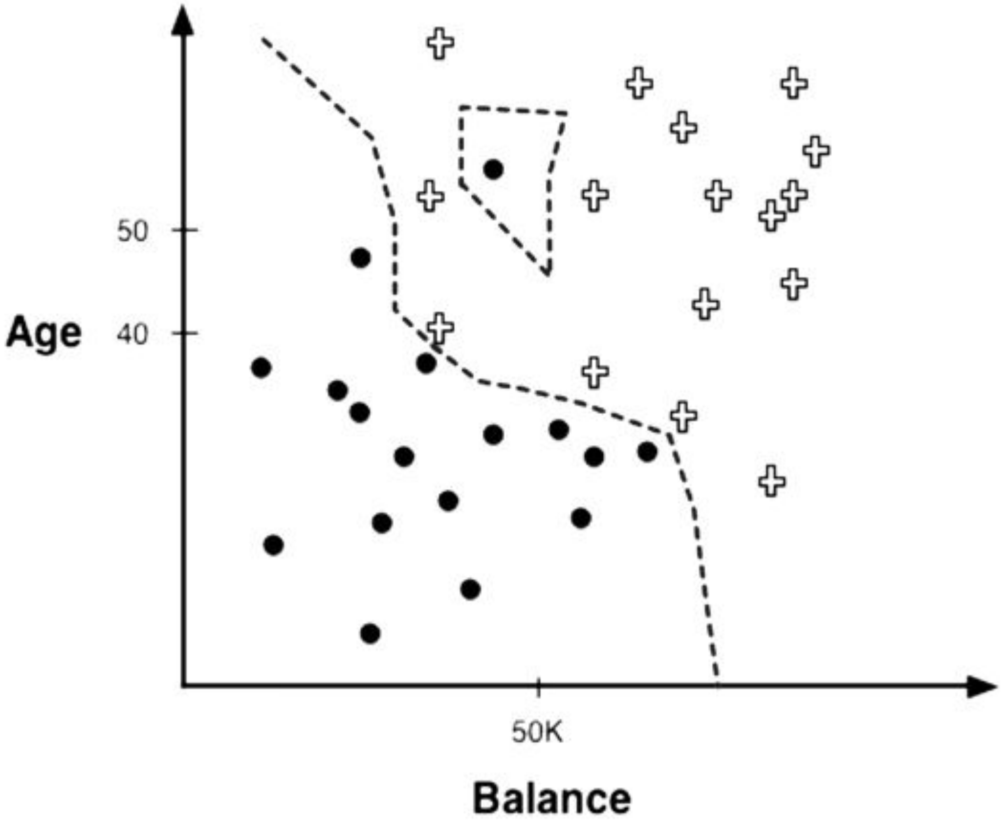
Source: DSB

K-Nearest Neighbors Classification

- Majority vote

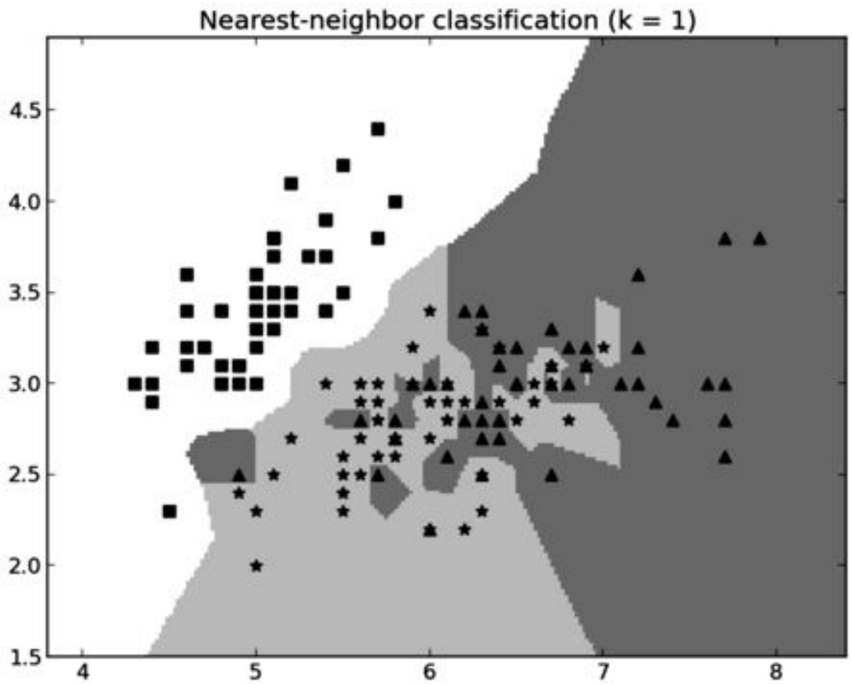
$$c(x) = \arg\max_{c} \sum_{y \in KNN(x)} \left[class(y) = c \right]$$

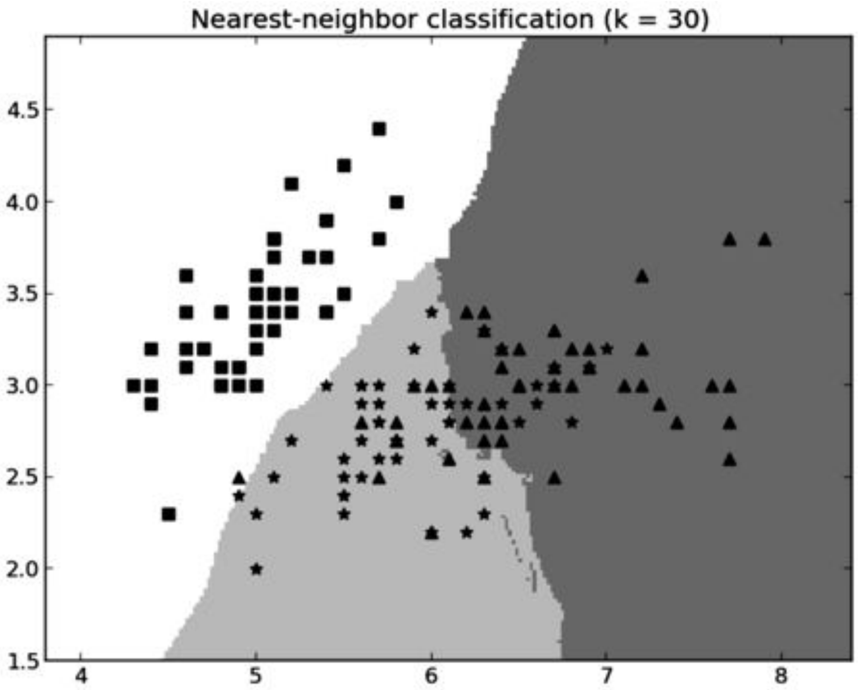
K-Nearest Neighbors



Source: DSB

K-Nearest Neighbors





Source: DSB

K-Nearest Neighbors Classification

- Weighted by similarity

$$w(x,y) = \frac{1}{dist(x,y)^2}$$

Name	Distance	Similarity weight	Contribution	Class
Rachael	15.0	0.004444	0.344	No
John	15.2	0.004348	0.336	Yes
Norah	15.7	0.004032	0.312	Yes
Jefferson	122.0	0.000067	0.005	No
Ruth	152.2	0.000043	0.003	No

$$p(c \mid x) = \frac{\sum_{y \in KNN(x)} w(x,y) \cdot [class(y) = c]}{\sum_{y \in KNN(x)} w(x,y)}$$

Source: DSB

K-Nearest Neighbors Regression

- Weighted by similarity

$$w(x,y) = \frac{1}{dist(x,y)^2}$$

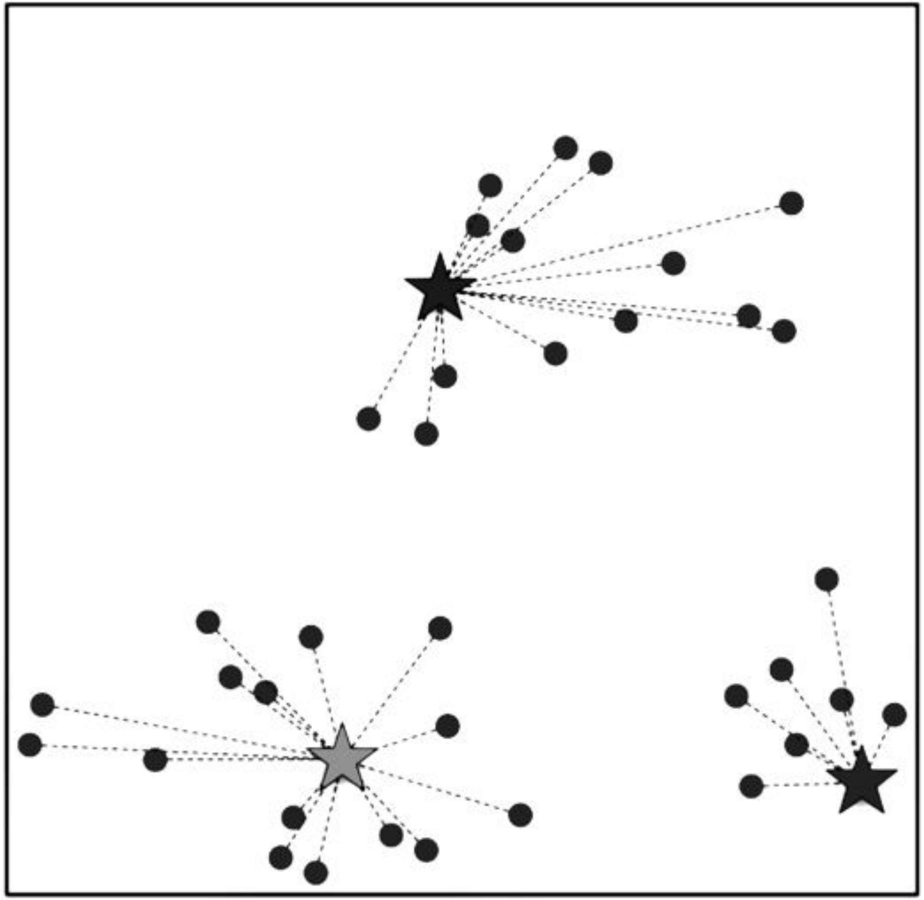
$$f(x) = \frac{\sum_{y \in KNN(x)} w(x,y) \cdot t(y)}{\sum_{y \in KNN(x)} w(x,y)}$$

Unsupervised Learning
Clustering
K-Means, Hierarchical

Clustering Goals

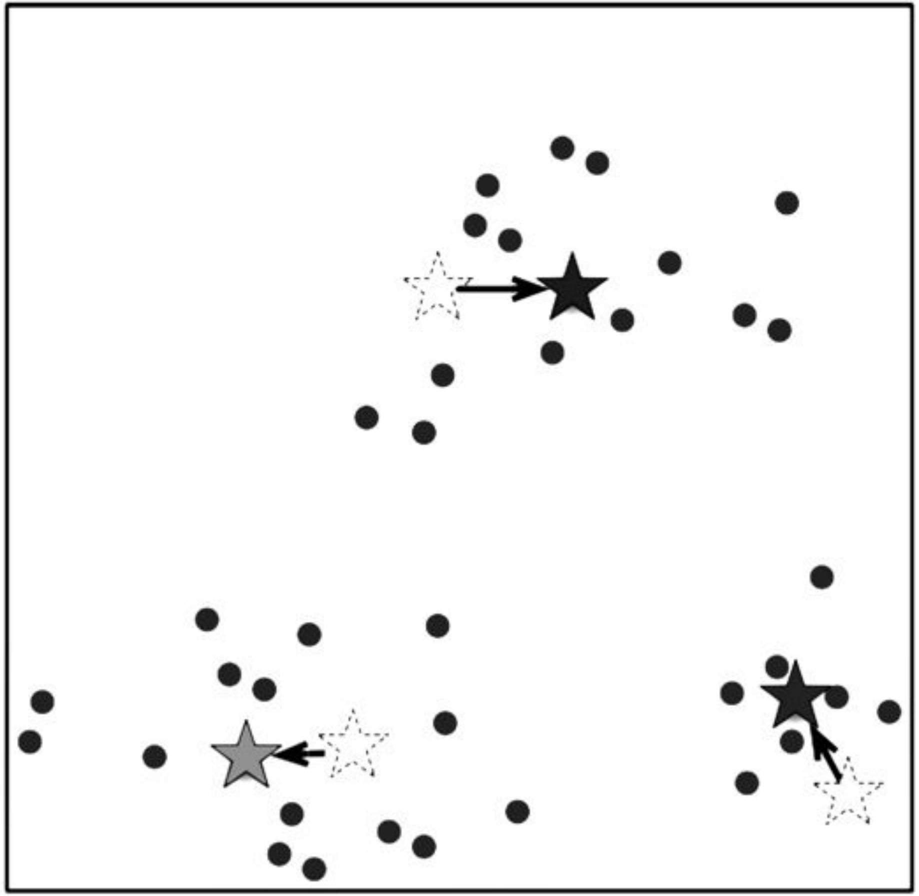
- Reliably achieve high accuracy across domains
- Handle high data dimensionality
- Scale to large datasets

K-Means Clustering



Source: DSB

K-Means Clustering

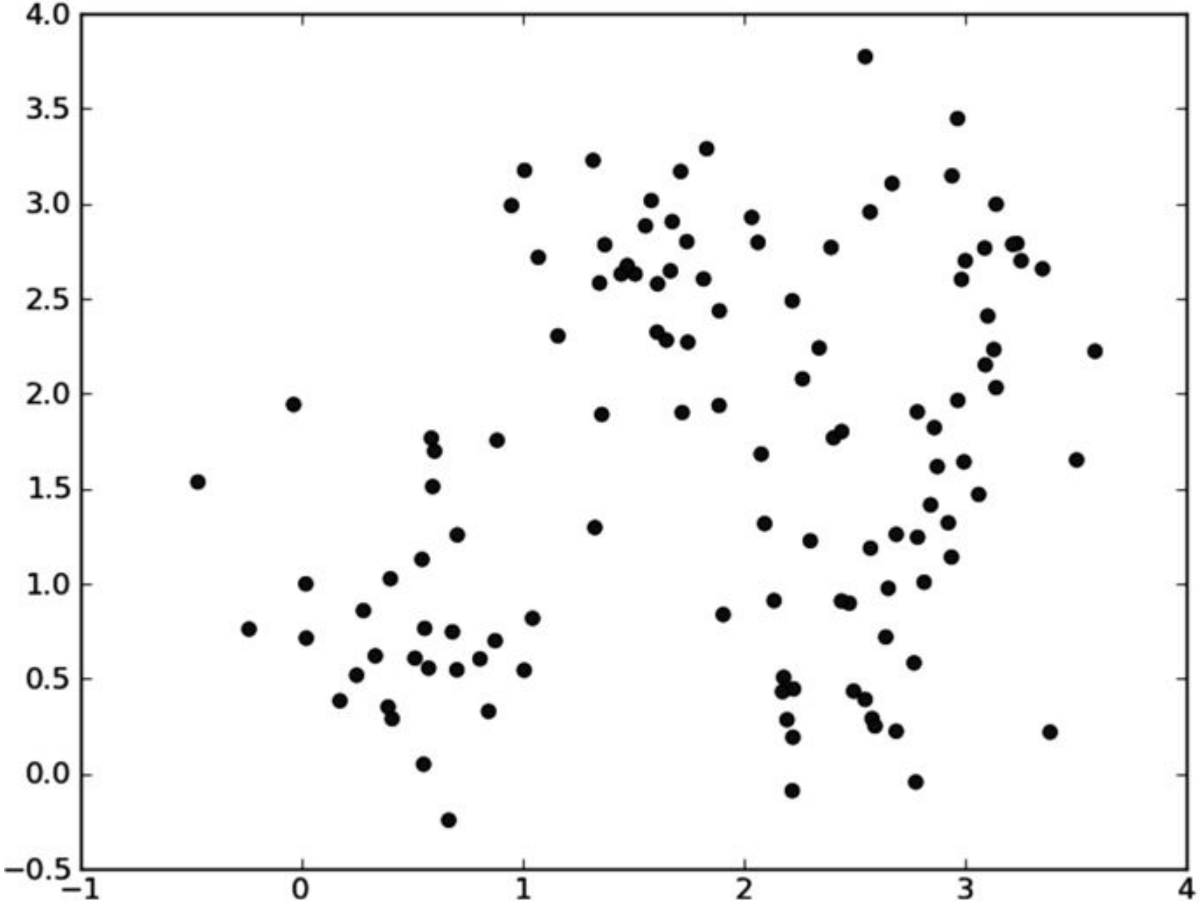


Source: DSB

K-Means Clustering

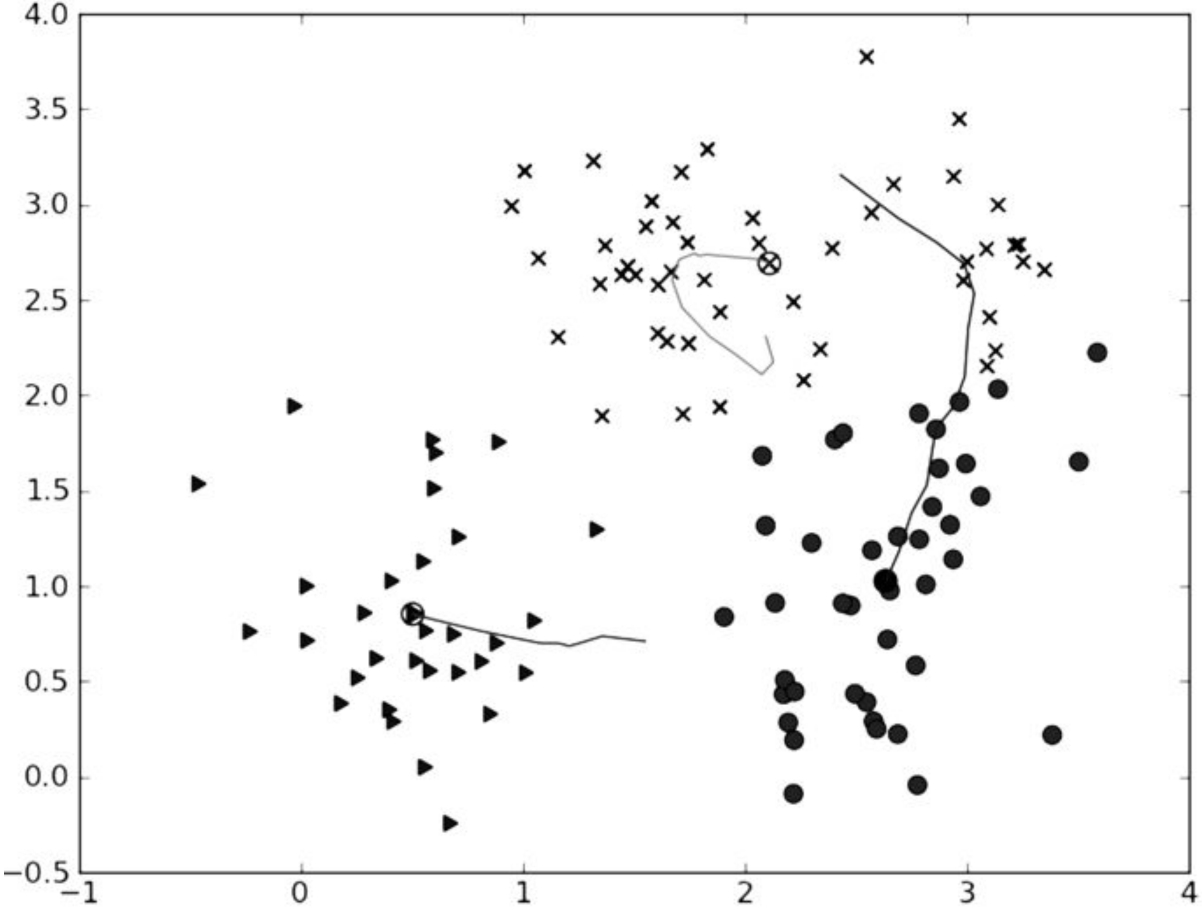
- Choose k
- Initialize cluster centroids as k random examples
- Repeat until convergence
  1. For each example: find its nearest cluster centroid and label example as belonging to that cluster.
  2. For each cluster centroid: update to mean of its labeled examples.

K-Means Clustering



Source: DSB

K-Means Clustering



Source: DSB

K-Means Clustering

- Objective

$$\min_{c(x_i)} \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \mu_{c(x_i)} \right)^2$$

K-Means Clustering

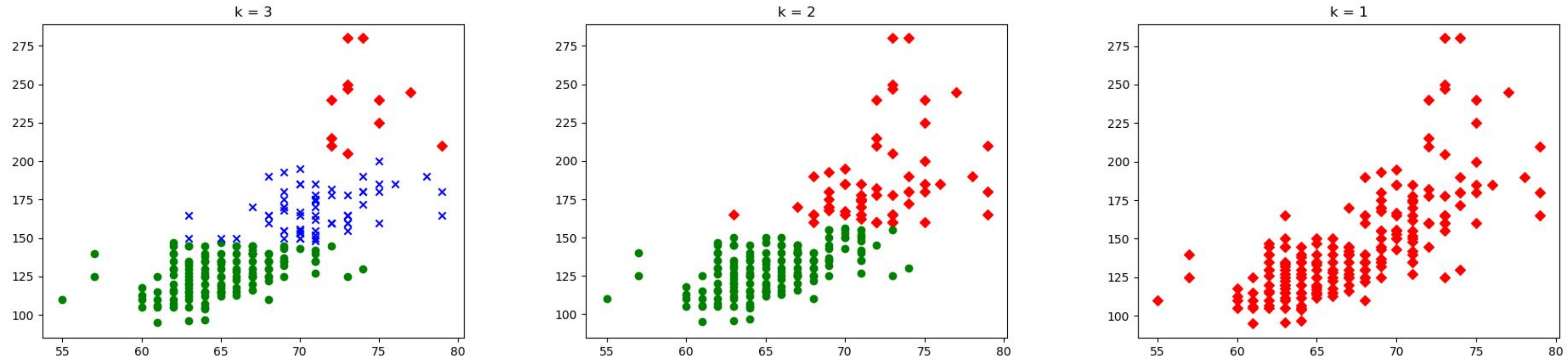
- Choose number of clusters k

- Choose number of experiments t
- For each experiment run K-means

- Select best clustering among t experiments minimizing objective

K-Means Clustering

- Choosing k

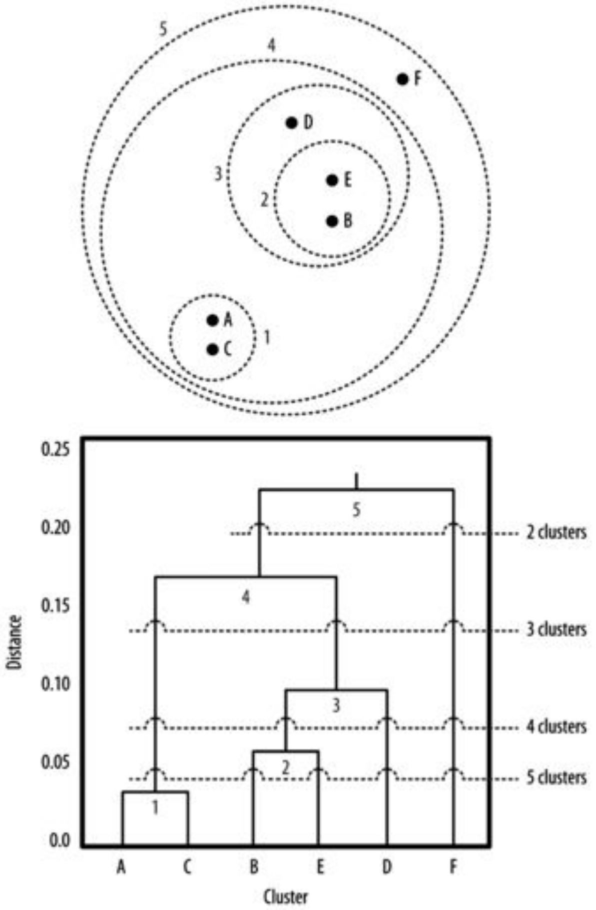


Hierarchical Clustering

- Bottom up: agglomerative
- Each sample starts in its own cluster, and pairs of clusters are merged

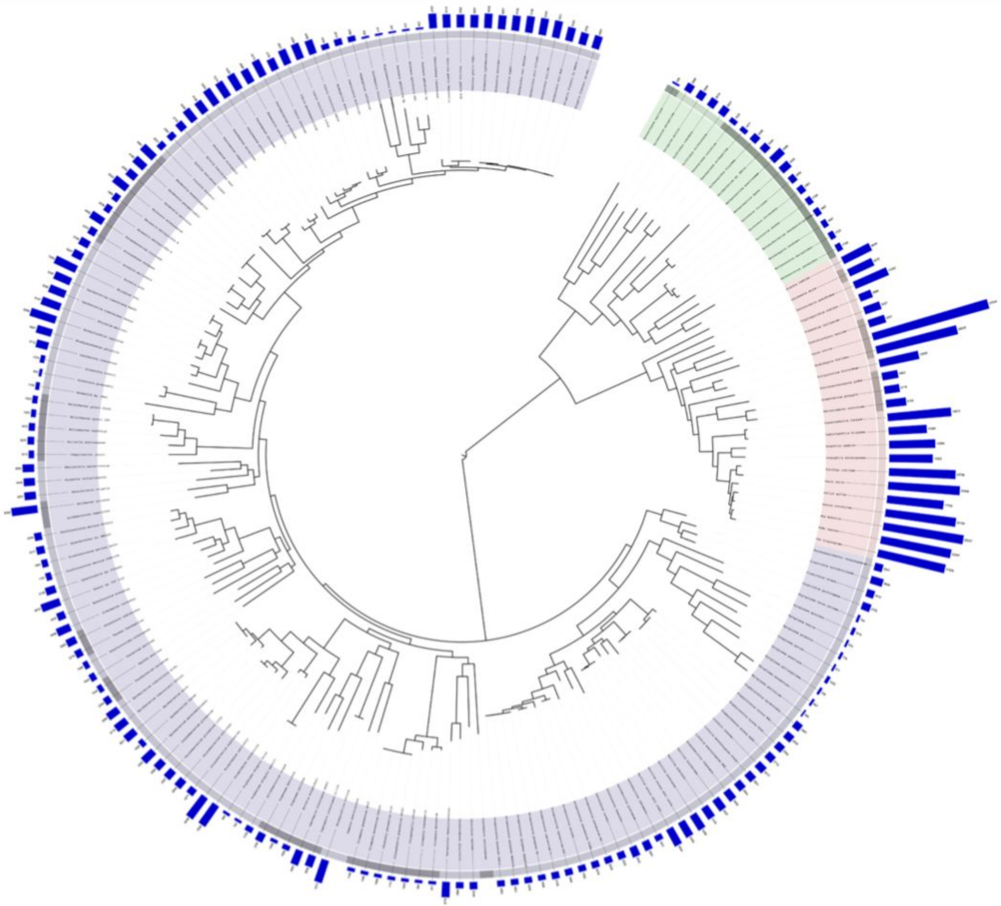
- Top down: divisive
- All samples start in one cluster, and splits are performed recursively

Hierarchical Clustering: Dendrogram



Source: DSB

Phylogenetic Tree of Life: Hierarchical clustering of species

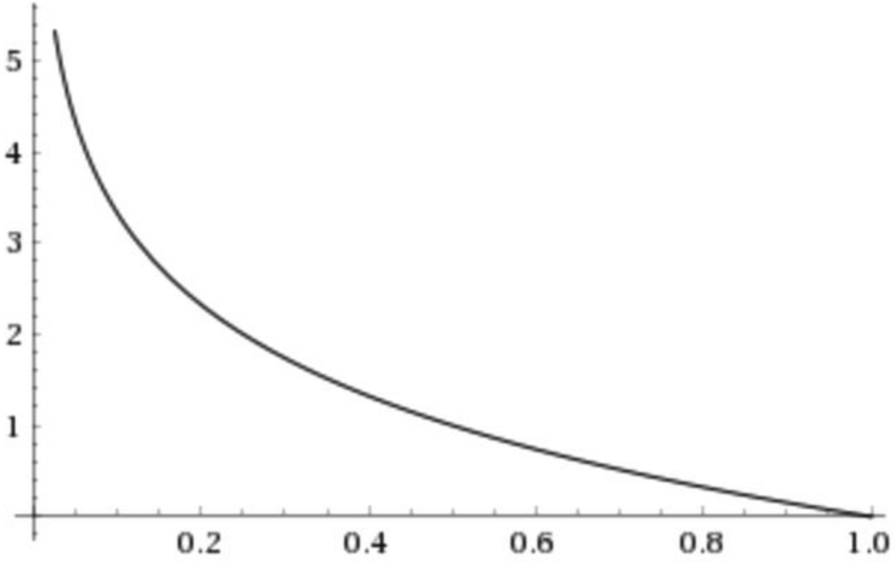


Source: DSB

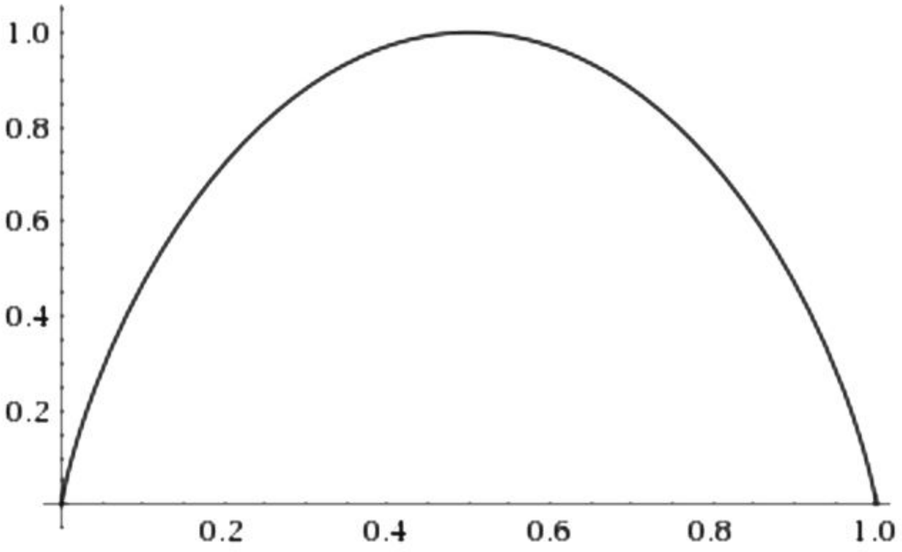
Clustering Quality

- Normalized mutual information
- Adjusted mutual information
- Rand index

Information and Entropy



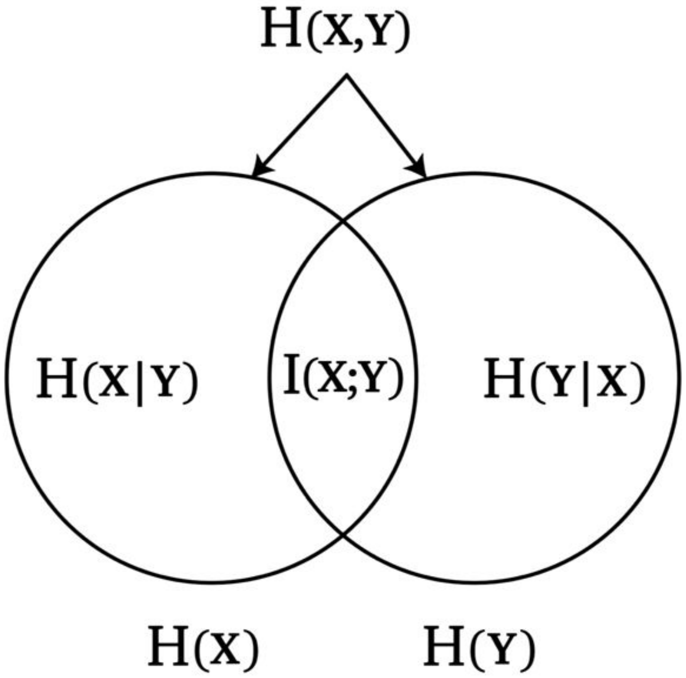
Information: $-\log_2(p)$ where $p \in [0, 1]$



Entropy: $-p \log_2(p) - (1-p)\log_2(1-p)$ where $p \in [0, 1]$

Mutual Information

- Reduction in uncertainty of X due to knowledge of Y
- $I(X;Y) = H(X) - H(X|Y)$

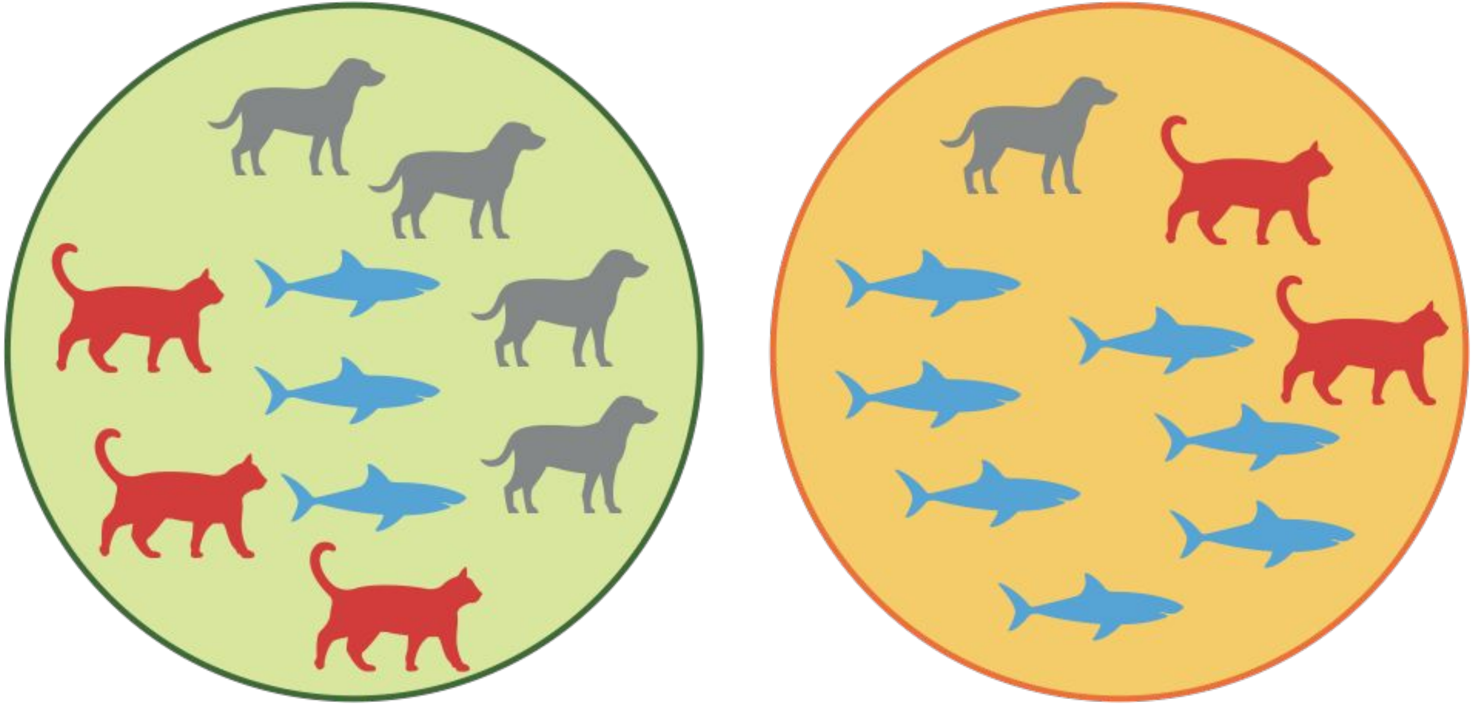


Clustering Quality

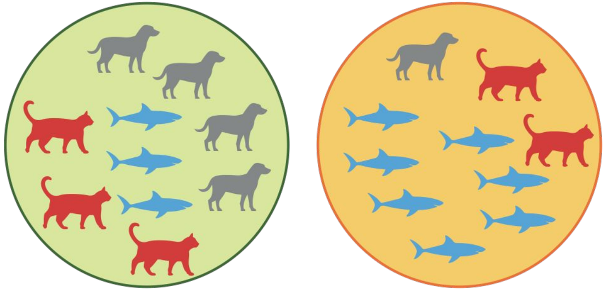
- Normalized mutual information
- Between class labels Y and cluster labels C

$$\frac{I(Y;C)}{\sqrt{H(Y)H(C)}}$$

Clustering Quality

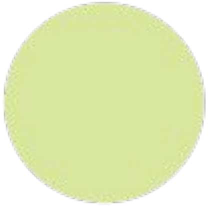


Clustering Quality



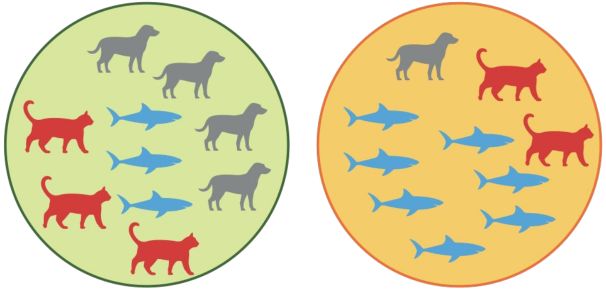
- Probability () = 5/20 = 1/4
- Probability () = 5/20 = 1/4
- Probability () = 10/20 = ½
- H(Y) = - 1/4log(1/4) - 1/4log(1/4) - 1/2log(1/2) = 3/2

Clustering Quality

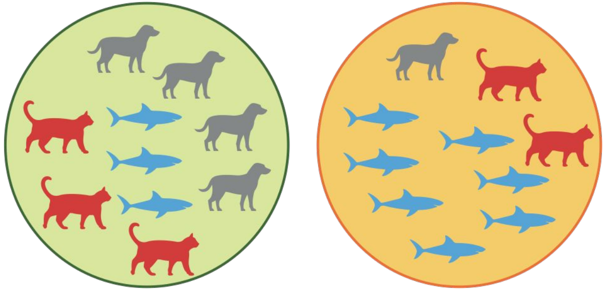
- Probability () = 10/20 = 1/2

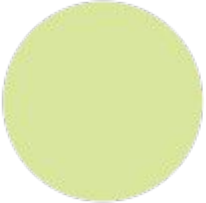
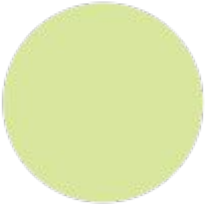
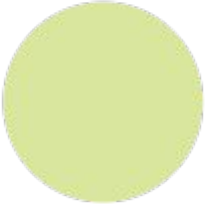
- Probability () = 10/20 = 1/2

- H(C) = - 1/2log(1/2) - 1/2log(1/2) = 1



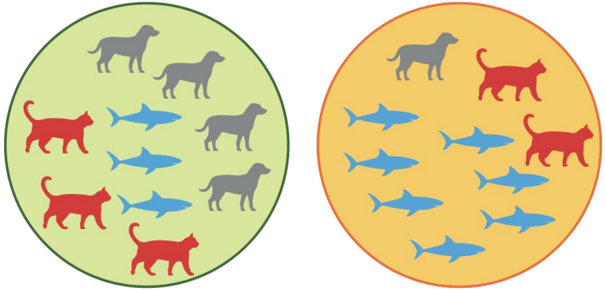
Clustering Quality

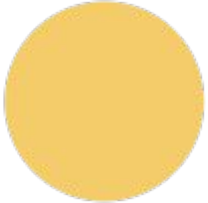
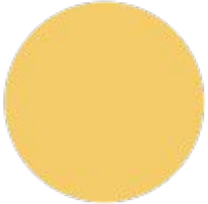


- Probability ( | ) = 3/10
- Probability ( | ) = 4/10 = 2/5
- Probability ( | ) = 3/10
- H(Y|C=1) = -P(C=1) sum_y P(Y=y|C=1)log(P(Y=y|C=1))

Clustering Quality

- Probability ( | ) = 2/10 = 1/5

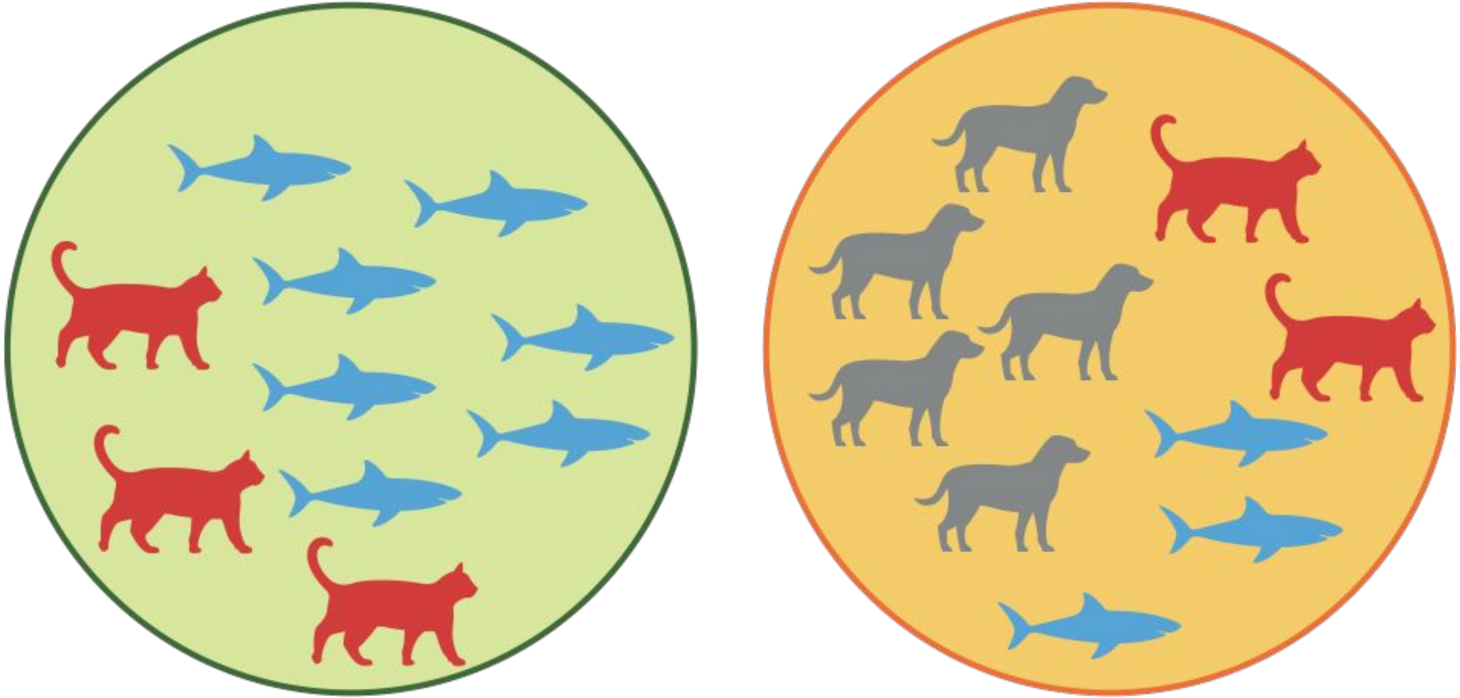


- Probability ( | ) = 1/10
- Probability ( | ) = 7/10
- H(Y|C=2) = -P(C=2) sum_y P(Y=y|C=2)log(P(Y=y|C=2))

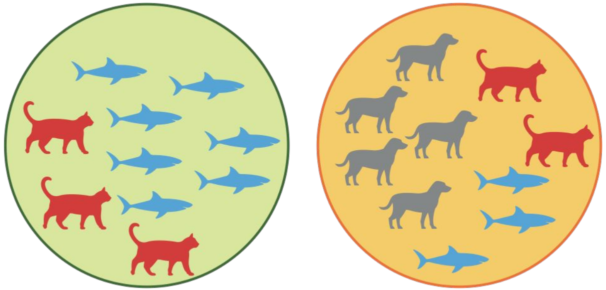
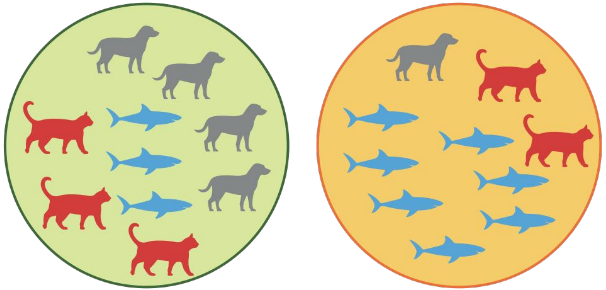
Clustering Quality

- $H(Y|C) = H(Y|C=1) + H(Y|C=2)$
- $I(Y;C) = H(Y) - H(Y|C)$

Clustering Quality



Clustering Quality

- NMI() > NMI()

Clustering Explainability

- Cluster
- Classify based on cluster labels using tree model

Unsupervised Learning
Dimensionality Reduction
Principal Component Analysis

Dimensionality Reduction

- Storage
- Computation
- Visualization

Dimensionality Reduction

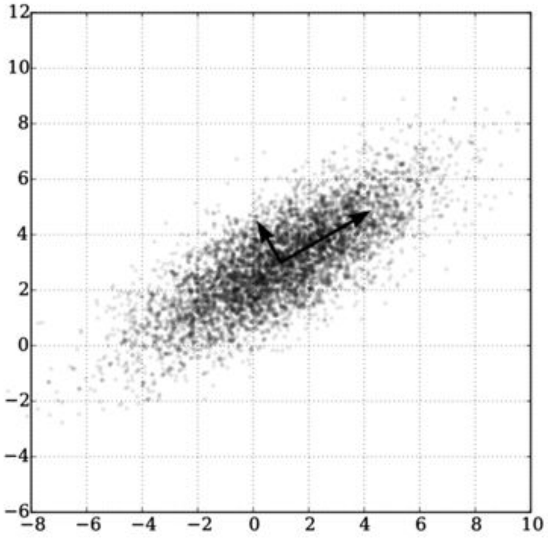
- Principal component analysis (PCA)
  - Maximize variance
  - Minimize mean squared error

- Find new basis in which vectors maximize variance

- Orthogonal linear transformation of data to new coordinate system such that greatest variance by projection of data is on 1st coordinate (first principal component), 2nd greatest variance is on 2nd coordinate, and so on

Dimensionality Reduction

- Subtract mean, so new mean is 0
- Variance is then $$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$



Principal Component Analysis (PCA)

- Compute covariance matrix $\Sigma = \frac{1}{m} M^T M$
- Compute eigenvectors of covariance matrix by singular value decomposition

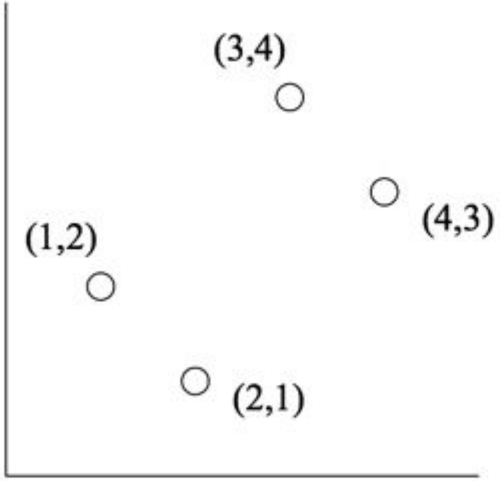
$$U, S, V^T = SVD(\Sigma)$$

$$\underset{mxk}{Z} = \underset{mxn}{M}\,\underset{nxk}{U_{1\cdots k}}$$

low dimensional representation of dimension k in rows of Z

examples in rows of M in dimension n
eigenvectors in columns of U

Principal Component Analysis (PCA)



$$M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}$$

Principal Component Analysis (PCA)

$$M^{\mathrm{T}}M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix}$$

$$(30 - \lambda)(30 - \lambda) - 28 \times 28 = 0$$

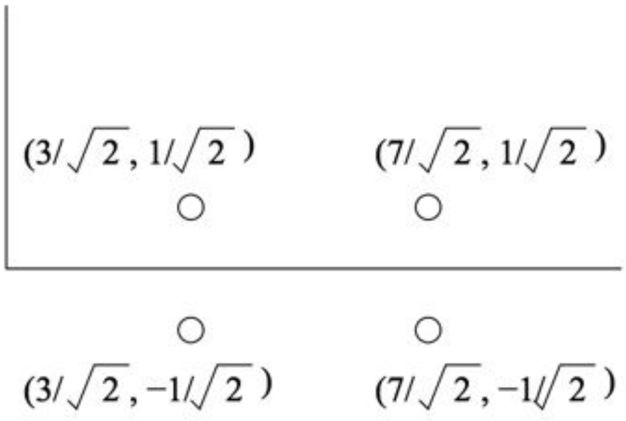
$$\lambda = 58 \text{ and } \lambda = 2$$

$$\begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 58 \begin{bmatrix} x \\ y \end{bmatrix} \qquad \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$$\begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 2 \begin{bmatrix} x \\ y \end{bmatrix} \qquad \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

$$E = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

Principal Component Analysis (PCA)



$$ME = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 3/\sqrt{2} & 1/\sqrt{2} \\ 3/\sqrt{2} & -1/\sqrt{2} \\ 7/\sqrt{2} & 1/\sqrt{2} \\ 7/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Principal Component Analysis (PCA)

E_k be the first k columns of E

ME_k is a k-dimensional representation of M

$$ME_1 \quad \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 3/\sqrt{2} \\ 3/\sqrt{2} \\ 7/\sqrt{2} \\ 7/\sqrt{2} \end{bmatrix}$$

Principal Component Analysis (PCA)

- Choose k such that

  99% of variance is retained $$\frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{n} S_{ii}} \geq 0.99$$

Principal Component Analysis (PCA)

- Computation speedup
- Dataset $\{x^{(i)}, y^{(i)}\}$
- Examples $\{x^{(1)}, \ldots, x^{(m)}\}$
- PCA of examples $\{z^{(1)}, \ldots, z^{(m)}\}$
- Train on low dimensional representation $\{z^{(i)}, y^{(i)}\}$

Dimensionality Reduction: t-SNE

- t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Map high-dimensional points to 2D/3D points such that:
  - Similar points map to nearby points
  - Dissimilar points map to distant points

http://projector.tensorflow.org

Dimensionality Reduction: t-SNE

- t-Distributed Stochastic Neighbor Embedding (t-SNE)
- pij measures similarity between xi and xj
- qij measures similarity between yi and yj
- Minimize Kullback Leibler divergence between p and q.
- Low dimensional map reflects similarities between high dimensional points.