

K6312 Information Mining & Analysis

Chen Zhenghua & Zhao Rui

Agenda

- Clustering
- K-means Algorithms
- Key Points Review

Preliminary: Supervised vs Unsupervised Learning

Supervised learning

Data: (x, y)

x is the data, y is label

Goal: Learn a function to map $x \rightarrow y$

Examples: classification, regression, object detection, etc.



→ Cat

classification

Preliminary: Supervised vs Unsupervised Learning

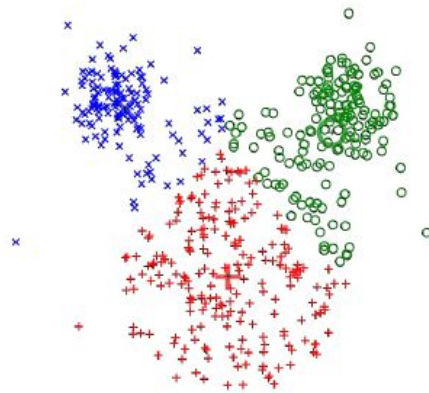
Unsupervised learning

Data: x

Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: clustering, feature learning, dimension reduction, etc.



clustering

Clustering

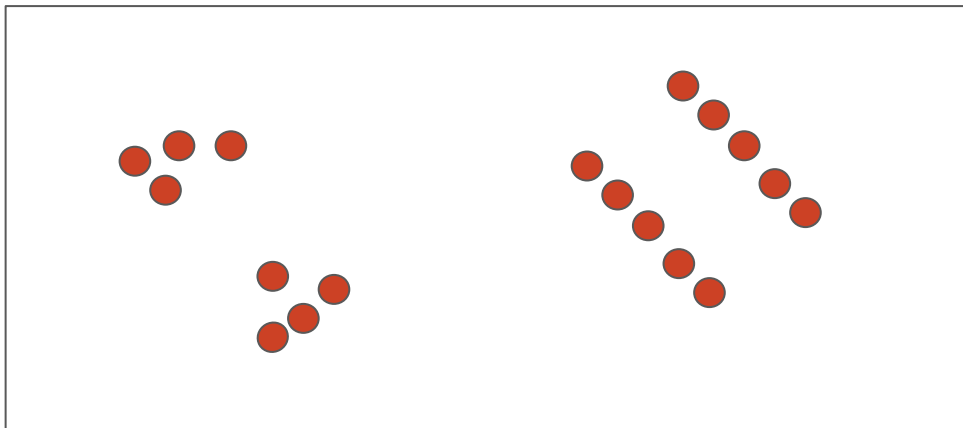
Clustering

- Unsupervised learning
- Requires data, but not labels
- Detect patterns
 - Group emails or search results
 - Customer shopping patterns
 - Regions of images
- Useful when do not know what you are looking for
- But: can get gibberish



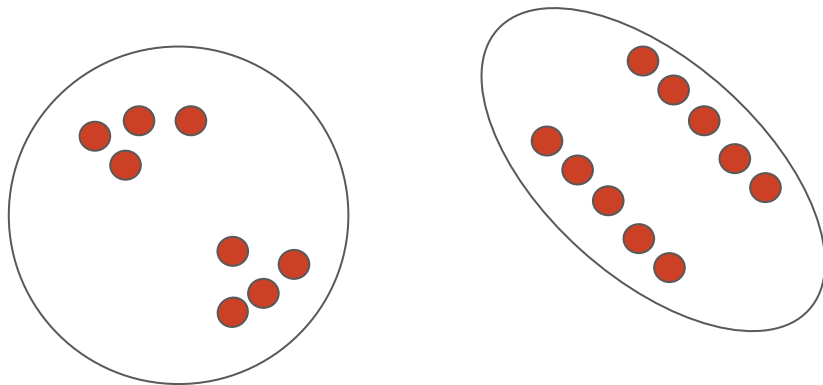
Clustering

- Basic idea: group similar instances together
- Examples: 2D points patterns



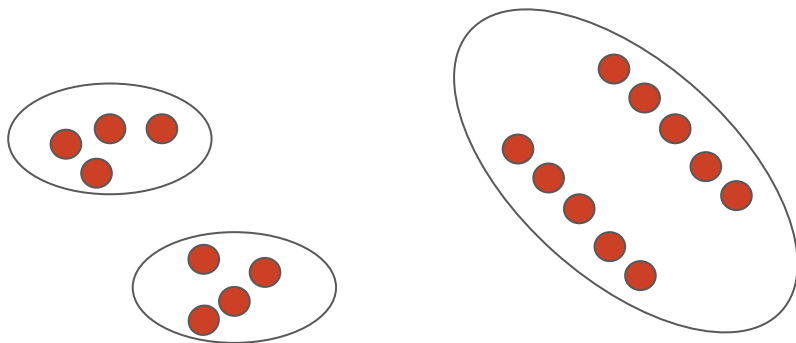
Clustering

- Basic idea: group similar instances together
- Examples: 2D points patterns



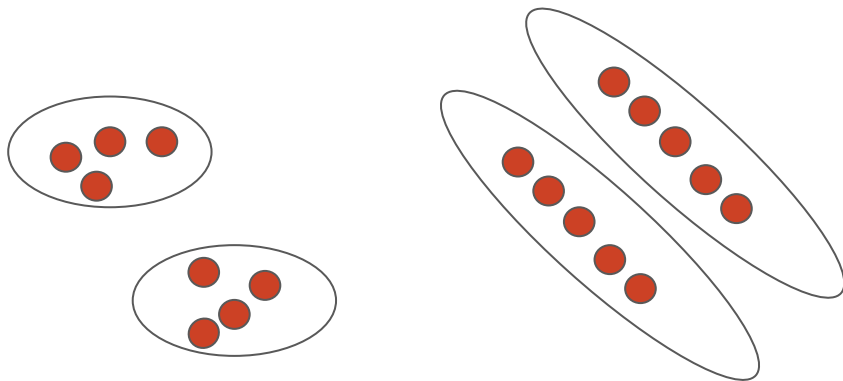
Clustering

- Basic idea: group similar instances together
- Examples: 2D points patterns



Clustering

- Basic idea: group similar instances together
- Examples: 2D points patterns



Similarity

- How to define similar

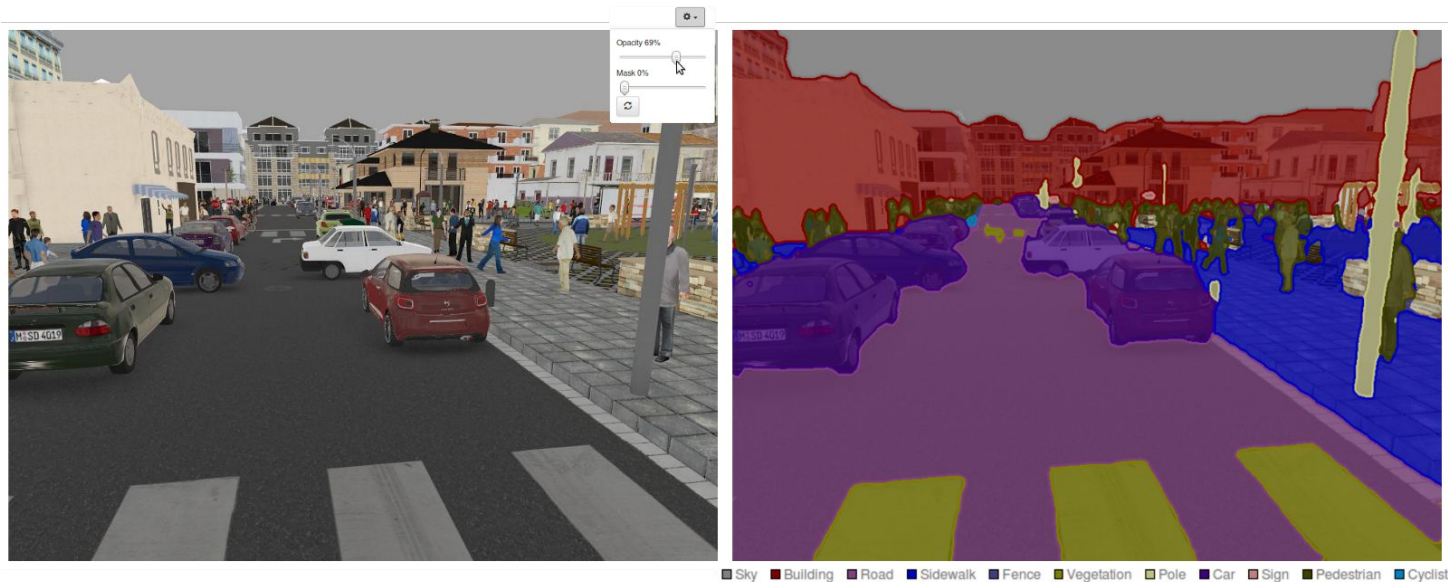
- The measures of similarity (or distance) between data samples are key components for clustering results
- One option: small Euclidean distance (squared)

$$dist(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||^2$$

- Similarity measures should match problem definition

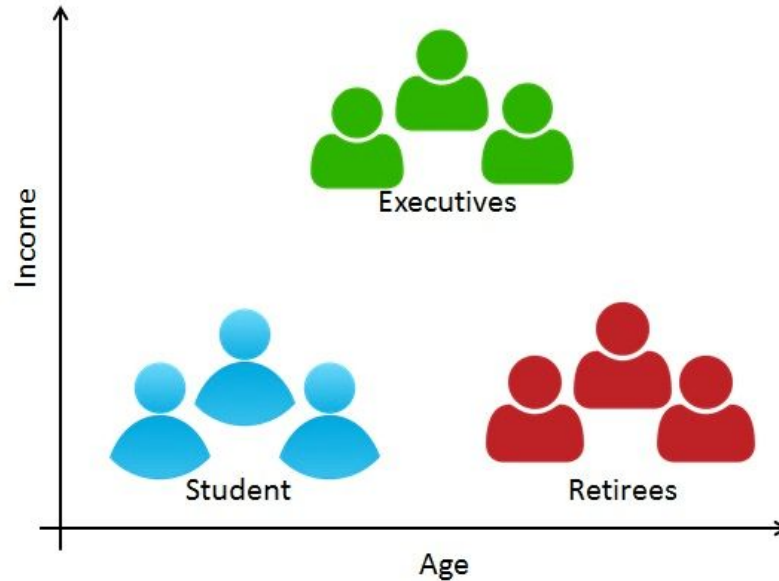
Clustering Applications

- Image Segmentation



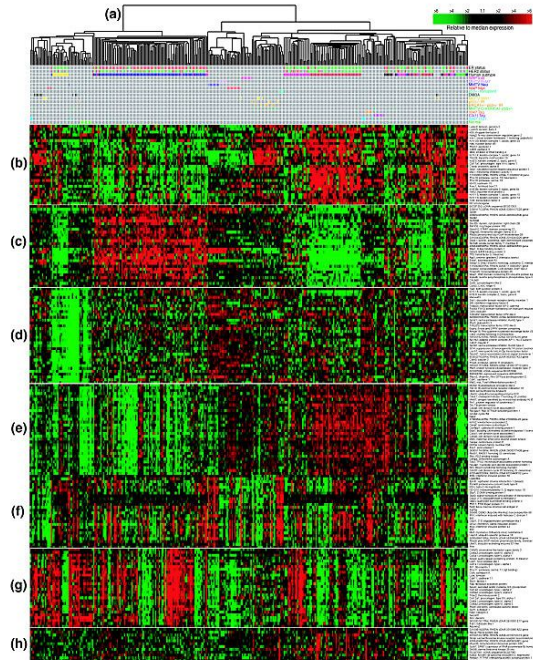
Clustering Applications

- Customer Segmentation



Clustering Applications

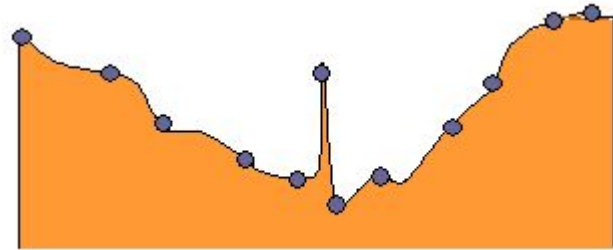
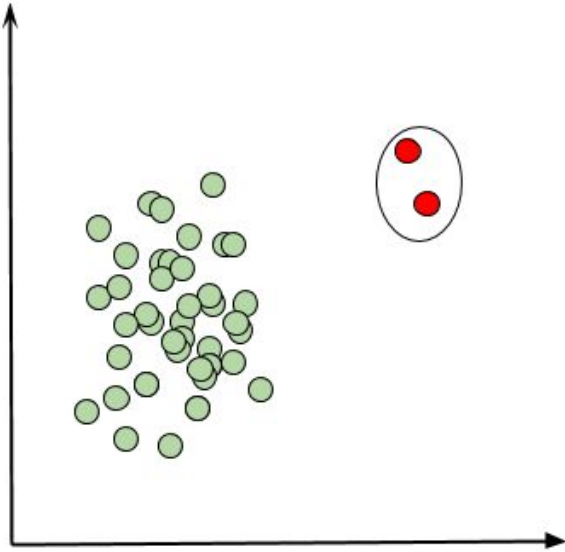
- Gene expression data clustering



source: Genome Biology 2007

Clustering Applications

- Anomaly detection



source: Floydhub

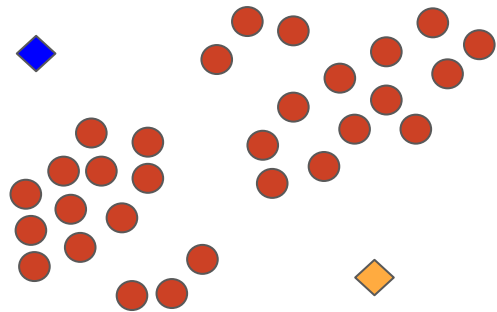
K-means

K-means

- An iterative clustering
 - Initialize: select K random points as cluster centers
 - Iteration process:
 - Calculate distance between data points and cluster centers
 - Assign data points to closet cluster center
 - Change the cluster center to the average of its assigned points
 - Stop when no points assignments change

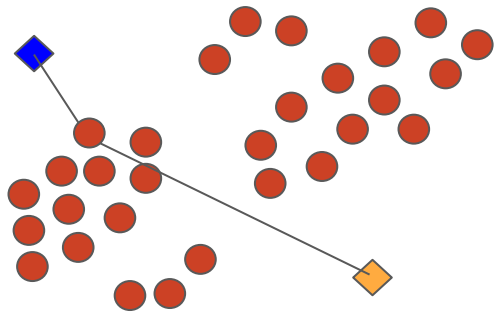
K-means clustering: examples

- Initialize 2 random points as cluster centers



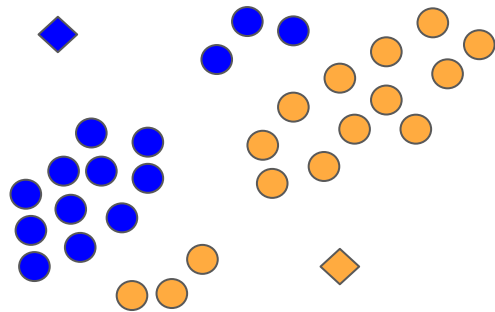
K-means clustering: examples

- Calculate distance (similarity measure) to each cluster centers



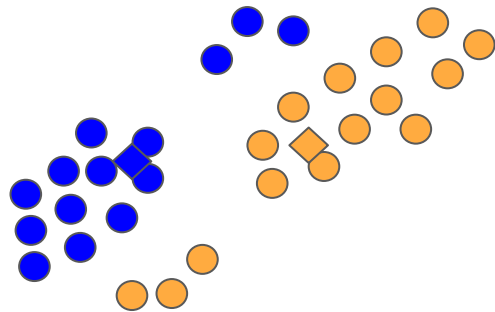
K-means clustering: examples

- Iteration one: Assign data points to closest cluster center



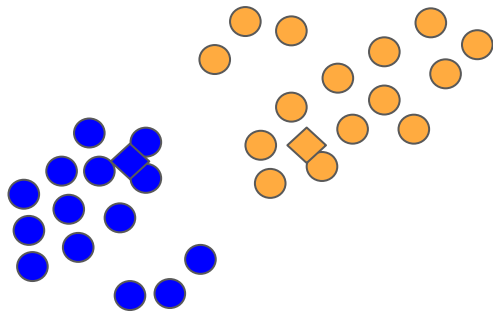
K-means clustering: examples

- Iteration one: Update the cluster center



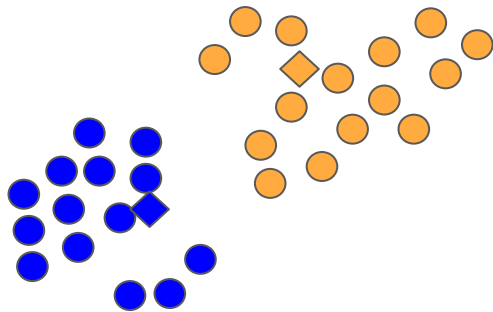
K-means clustering: examples

- Iteration two: Assign data points to closest cluster center



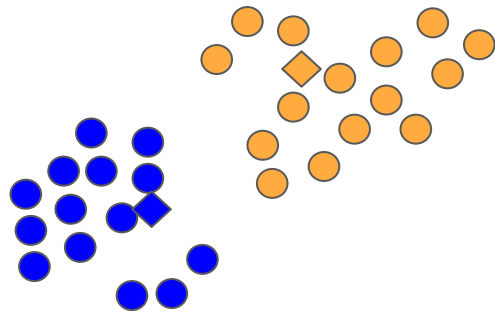
K-means clustering: examples

- Iteration two: Update the cluster center



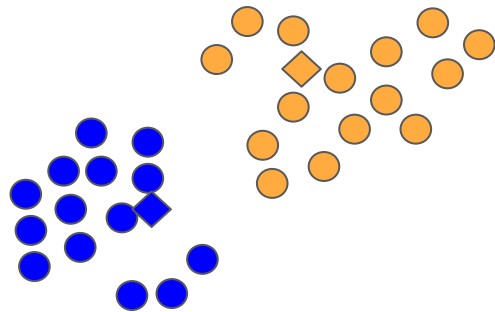
K-means clustering: examples

- Repeat until convergence



K-means clustering: examples

- Repeat until convergence

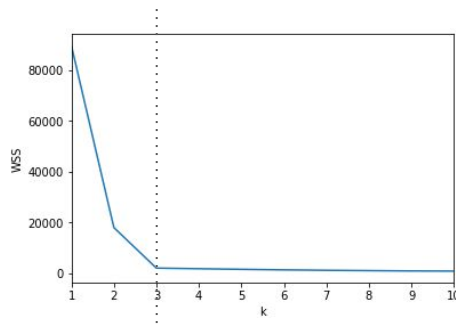
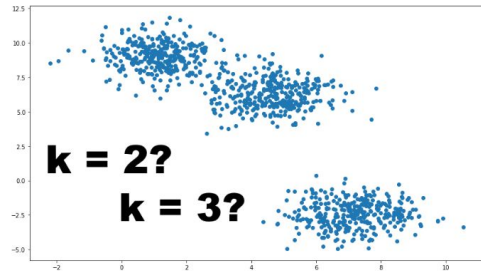


Stopping criteria

- How to define convergence?
 - No data points change clusters
 - Sum of the distances is minimized
 - Some maximum number of iterations is reached
- This algorithm is guaranteed to converge to a result (but maybe a local optimum)

How to find K?

- The number of cluster should be pre-defined
- One of the metrics can be the mean distance between data points and their cluster centroids
 - Draw the figure with the mean distance and the number of centroids
 - Elbow point: ***Within-Cluster-Sum of Squared Errors (WSS)***



Key points review

Lecture 1 & 2

- Basic concepts
 - What is data mining
 - The process of data mining
 - Some applications for classification/regression/clustering
- Basic for Python programming

Lecture 3 & 4

- Linear Regression

- Mean square error (MSE)

$$J(w_0, w_1) = \frac{1}{m} \sum_{i=1}^m [h(x_i) - y_i]^2$$

$$h(x) = w_0 + w_1x$$

$$h(\vec{x}) = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n$$

- Logistic Regression

- Entropy Loss

$$Loss(y, \tilde{y}) = -[y \log \tilde{y} + (1 - y) \log(1 - \tilde{y})]$$

- Evaluation of Classification

- Confusion matrix, precision, recall, F1

$$f(x) = \frac{1}{1 + e^{-(w*x+b)}}$$

Lecture 5 & 6

- Decision Tree
 - ID3: Information Gain (Entropy, continuous attributes, regression tasks)
 - Pros & Cons
 - Overfitting
- Ensemble Learning
 - Procedures of Bagging, Boosting, Stacking and Cascading

Lecture 7 & 8

- Neural Networks
 - Understand activation functions
 - Forward calculation
 - Training procedure
- Deep Learning
 - Properties of deep learning
 - Applications
 - Limitations

Lecture 9 & 10

- Support Vector Machine
 - How to use kernels
 - Merits and limitations
- Interpretable and Responsible Machine Learning
 - Application scenarios