JOHNS HOPKINS
CAREY BUSINESS SCHOOL

Lecture 6

BU.330.775 Machine Learning

Minghong Xu, PhD.
Associate Professor

Review

- Dimensions! Dimensions! Dimensions!
- PCA: projection method
- t-SNE: manifold method
- Use cases: visualization and feature extraction

Today’s Agenda

- Clustering and business usages
- Hands-on using MNIST
- Competition

Unsupervised Learning (Recap)

- Dimensionality reduction
  - Visualization
  - Factor analysis (Finance)
  - Natural language processing
  - Gene sequencing
- **Clustering**
  - **Product recommendations**
  - **Customer segmentation**
  - **Targeted marketing**
  - **Medical diagnostics**
- Association Rule (in Cloud Computing course)

Clustering

- Organize data into clusters such that
  - High intra-cluster similarity
  - Low inter-cluster similarity
- What is “similarity”?
  - Visual/appearance, …
  - Defined using distance, or correlation, etc.

Credit: Dr. Eric Xing, Introduction to Machine Learning
Carnegie Mellon University

Clustering vs Classification

- Like classification: each instance assigned to a group
- Unlike classification: an unsupervised task

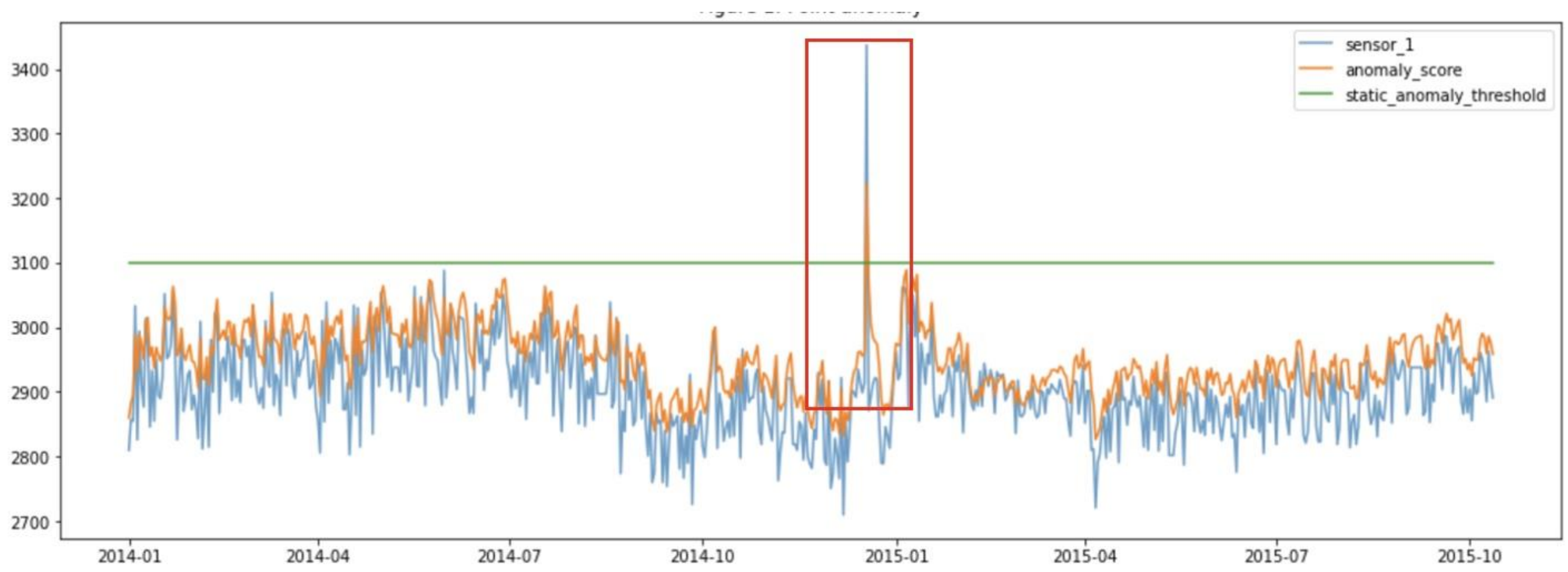
- *When to use classification? When to use clustering?*

labeled data

No labeled data,
also cases where labeling
can be expensive

Anomaly Detection

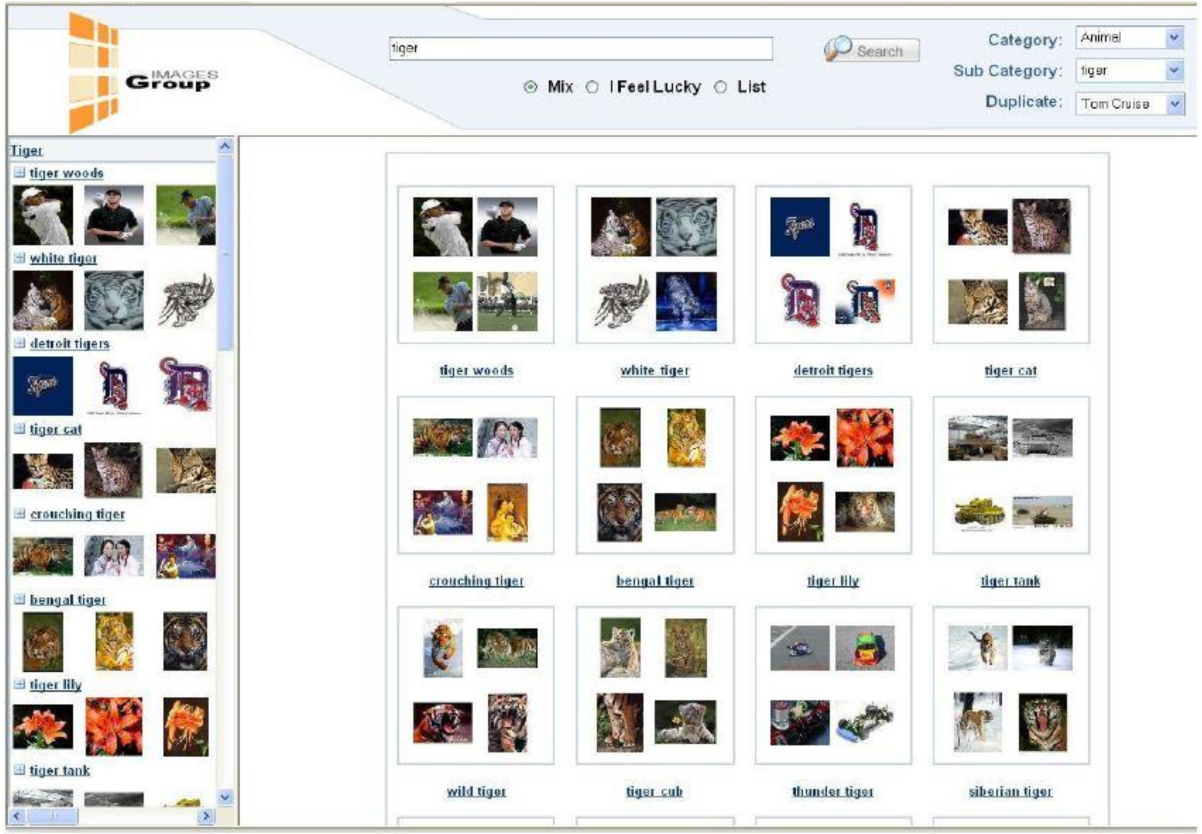
- Outlier detection
- Any instance having a low affinity to all clusters is likely to be an anomaly
- E.g., unusual number of requests per second



https://developer.ibm.com/learningpaths/get-started-anomaly-detection-api/what-is-anomaly-detection/

Search Engines

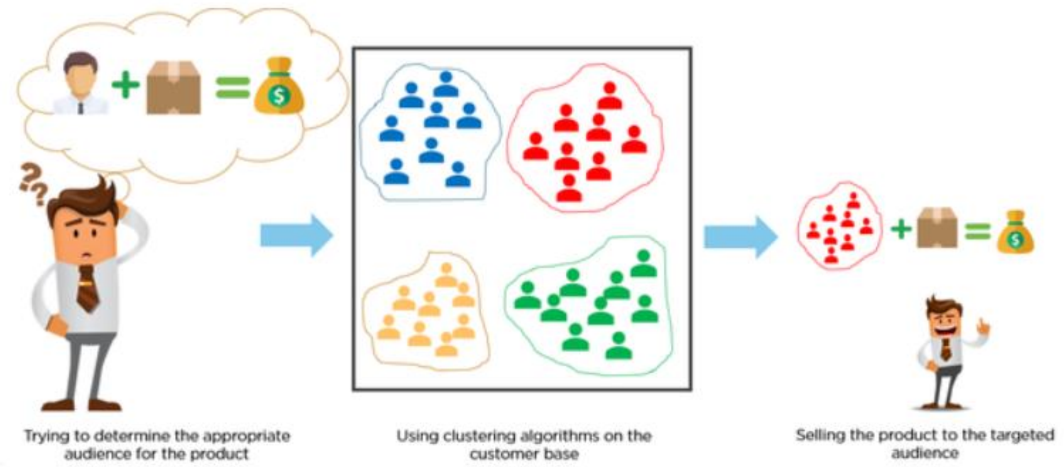
- Search for images that are similar to a reference image
- Apply clustering to all images
- Return images from the same cluster



https://www.microsoft.com/en-us/research/project/igroup-web-image-search-results-clustering/

Customer Segmentation

- Cluster customers based on purchases and/or activities
- Better understand your customers, adapt campaigns to each segment
- Widely used in recommender systems (in cloud computing course)
- Not identifying a class



https://www.quora.com/What-is-clustering

Image Segmentation

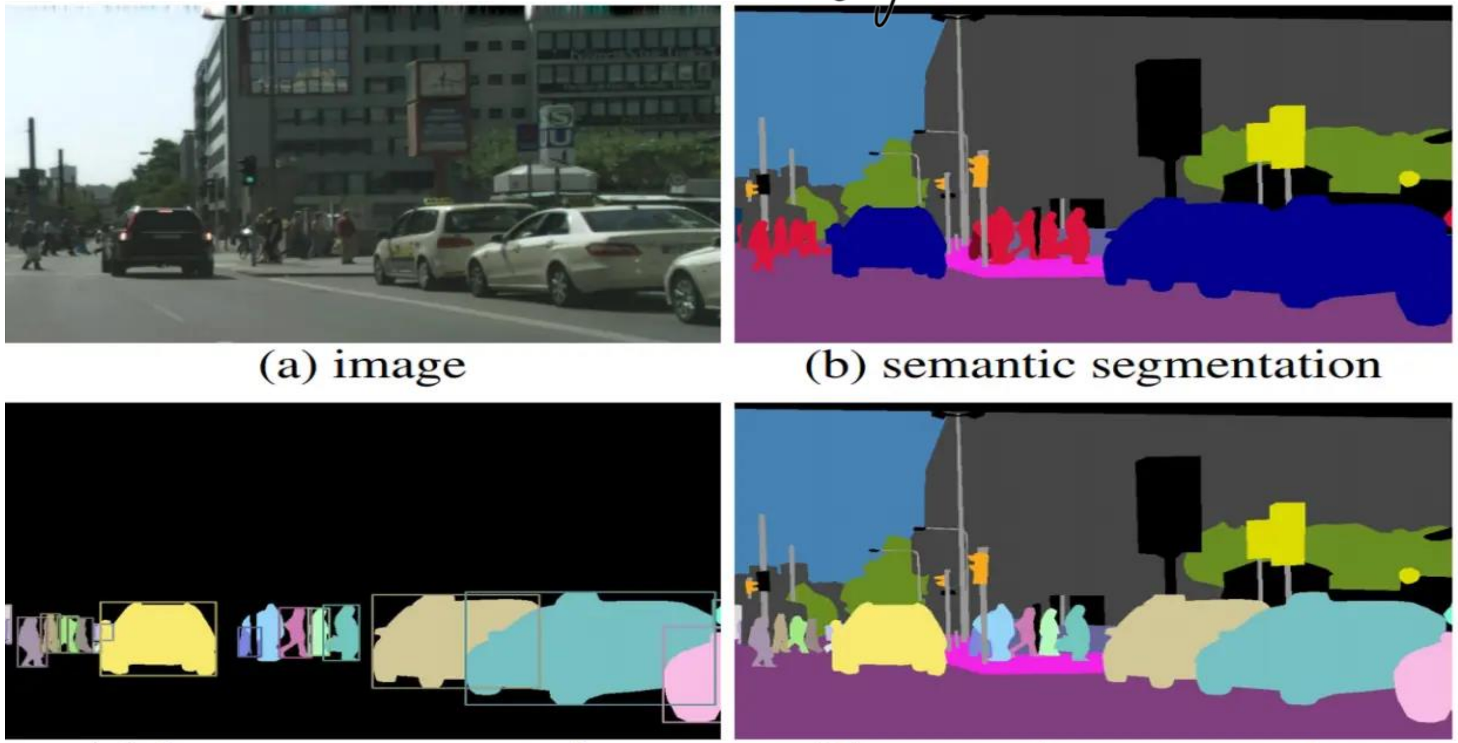
- Color segmentation: pixels with a similar color assigned to the same segment

Credit: James Hayes

Supervised Image Segmentation (Optional)

- Semantic segmentation: pixels belong to the same object type
  - E.g., a segment of all pedestrians
- Instance segmentation: pixels of the same individual object
  - E.g., different segment for each pedestrian

first colour segmentation
then object detection



(a) image (b) semantic segmentation

(c) instance segmentation (d) panoptic segmentation

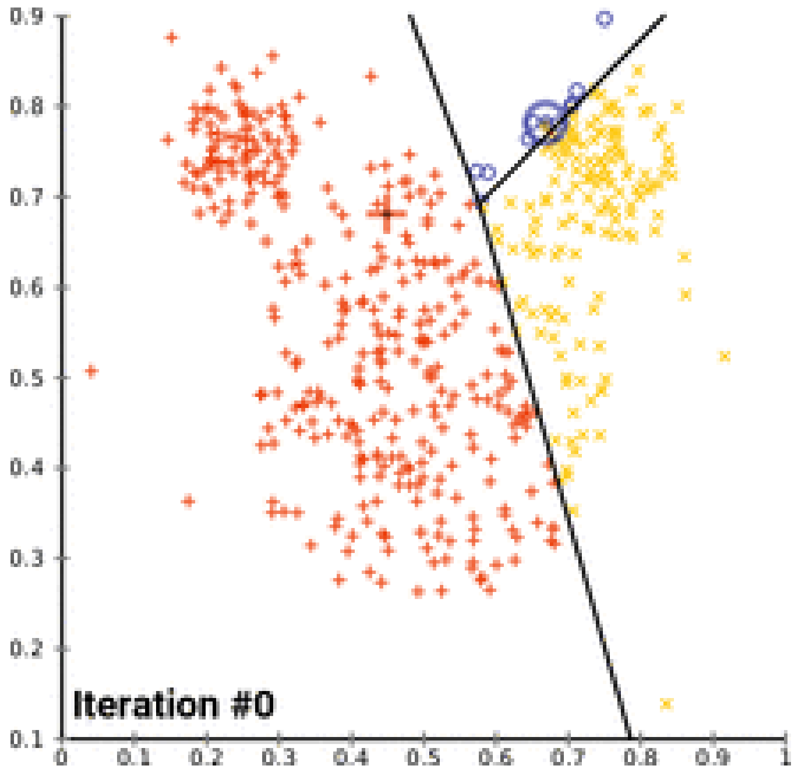
https://www.labellerr.com/blog/semantic-vs-instance-vs-panoptic-which-image-segmentation-technique-to-choose/

K-Means

- Partition *n* points into *k* clusters in which each point belongs to the cluster with the nearest mean

- *K* cluster centers or cluster **centroid**

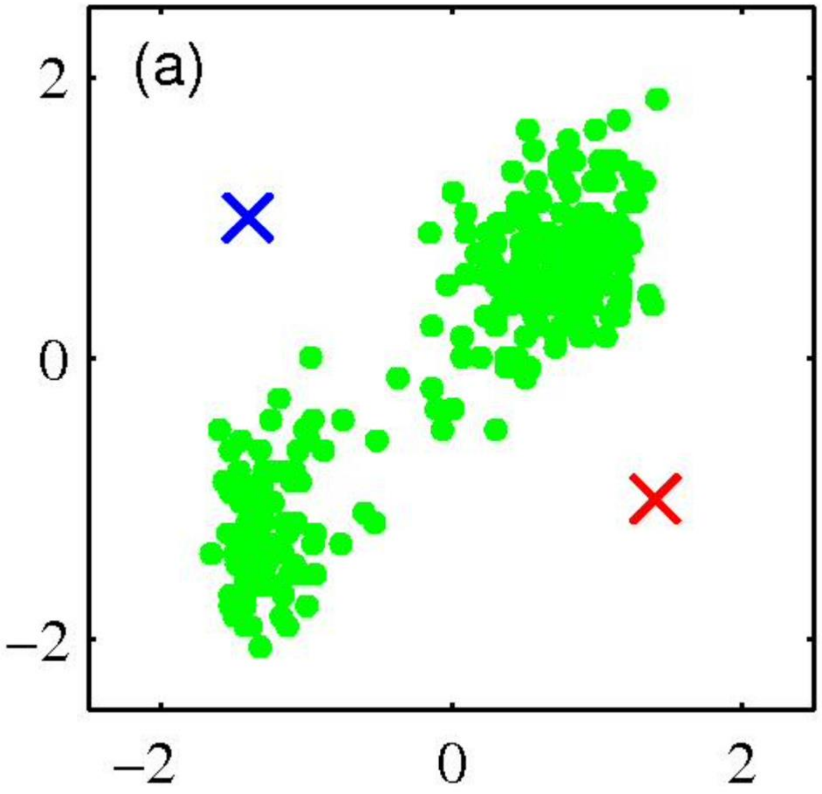
- Iterative algorithm



https://en.wikipedia.org/wiki/K-means_clustering

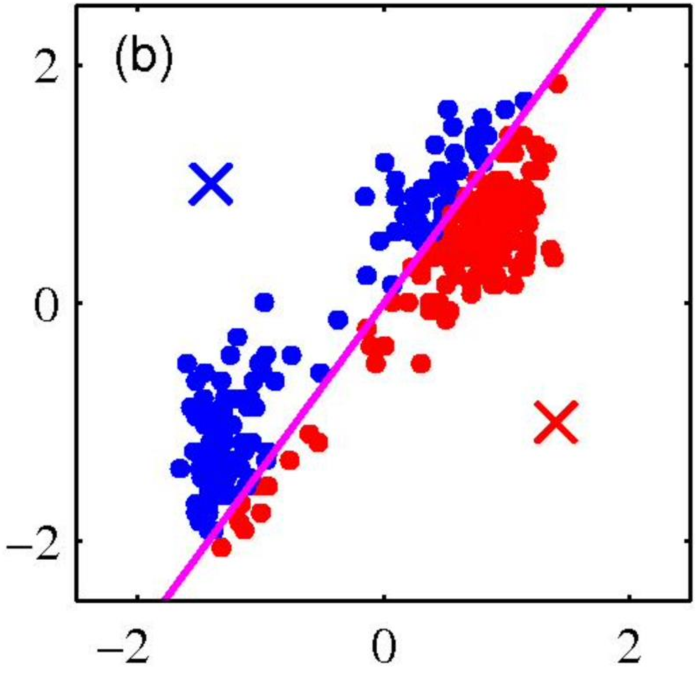
K-Means: Initialize

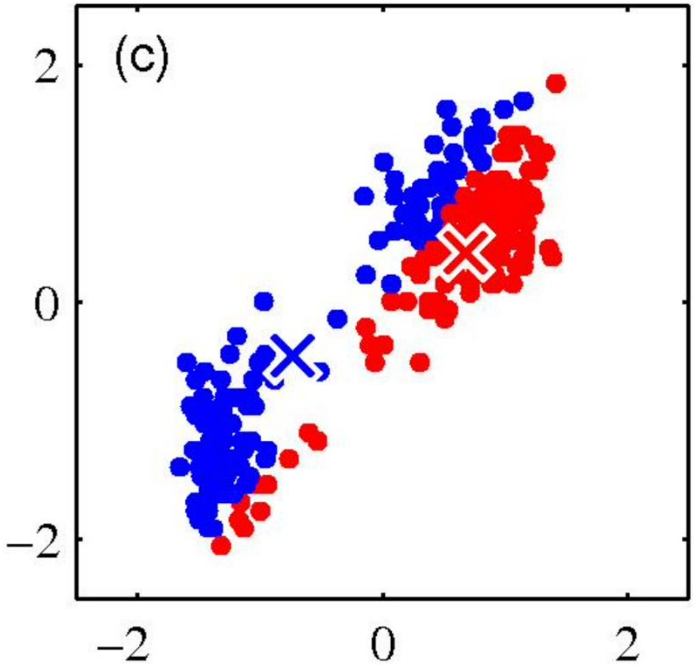
- Pick K random points as cluster centers



K-Means: Repeat

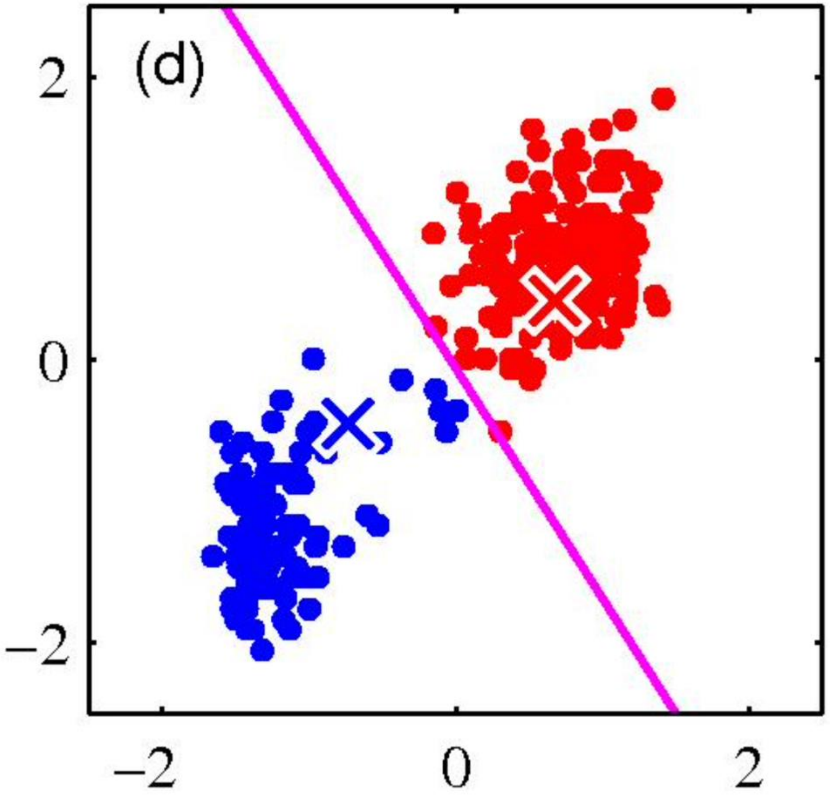
- Assign data points to closest cluster center
- Change the cluster center to the average of its assigned points





K-Means: Converge

- No cluster assignments change

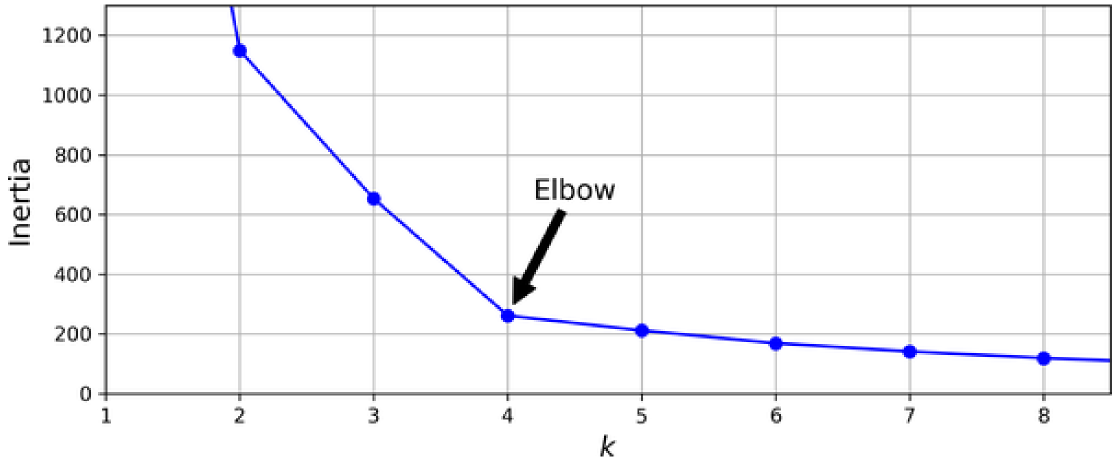


Performance Measures

Notes

- Inertia: sum of squared distances between the instances and their closest centroids
  - The lower the better. *Why?*
  - Generally decrease if k increases

```
>>> kmeans.inertia_
211.59853725816836
```



- There is another internal measure, silhouette score, not required
- External: compare to the true label

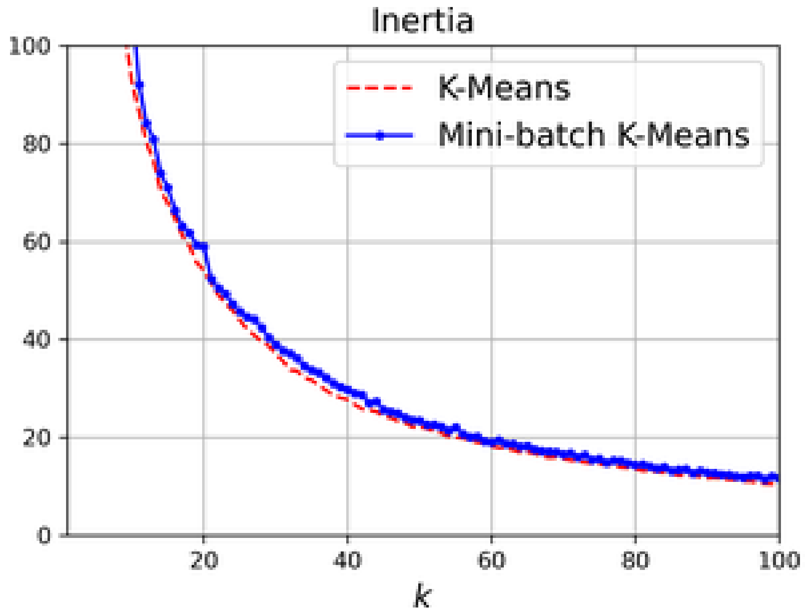
Hard Clustering vs Soft Clustering

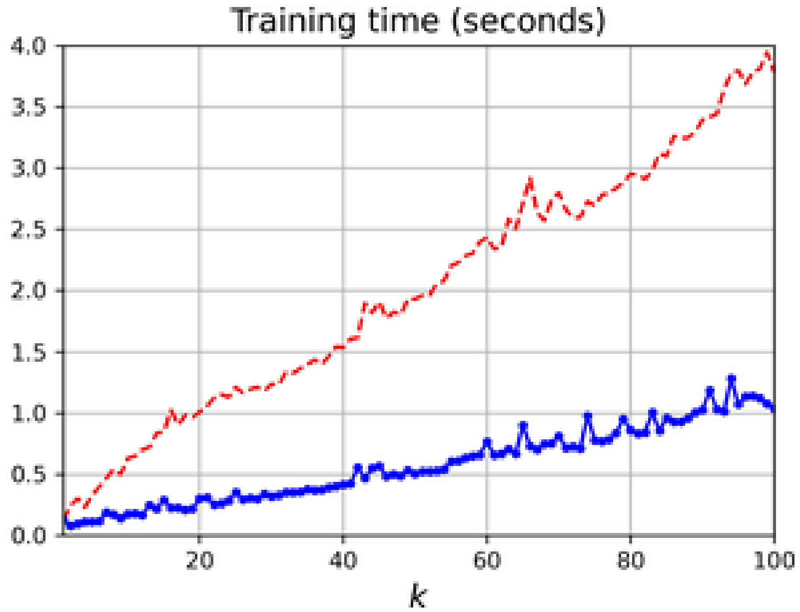
- Hard clustering
  - Assign each instance to a single cluster
- Soft clustering
  - Give each instance a score core per cluster
  - Score: distance/similarity between instance and the centroid

```
>>> kmeans.transform(X_new).round(2)
array([[2.81, 0.33, 2.9 , 1.49, 2.89],
       [5.81, 2.8 , 5.85, 4.48, 5.84],
       [1.21, 3.29, 0.29, 1.69, 1.71],
       [0.73, 3.22, 0.36, 1.55, 1.22]])
```

Mini-batch K-means

- Use mini-batches to update the centroids just slightly at each iteration
  - Instead of the full dataset
- Speed up the algorithm, especially when k is large

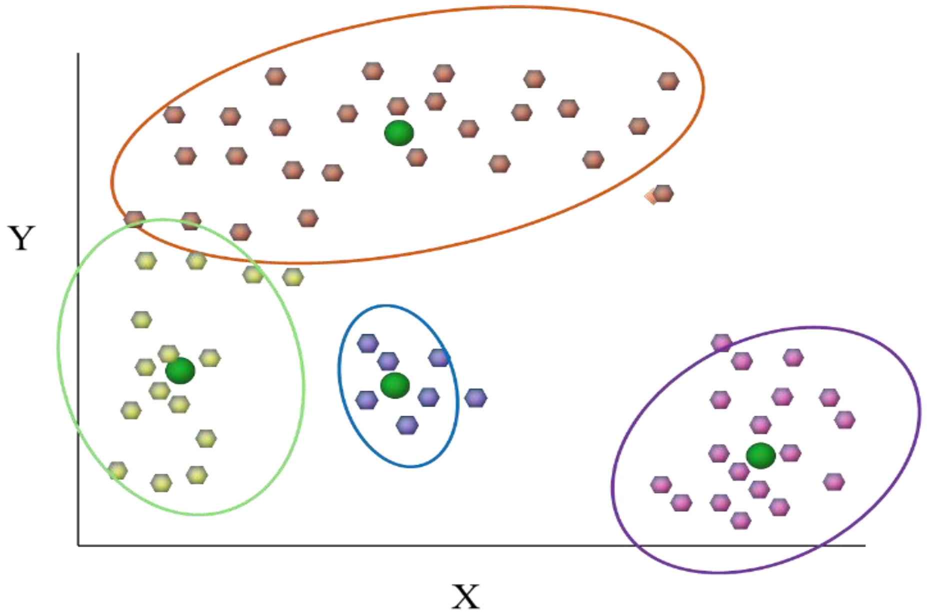




Issues of Clustering

Noten

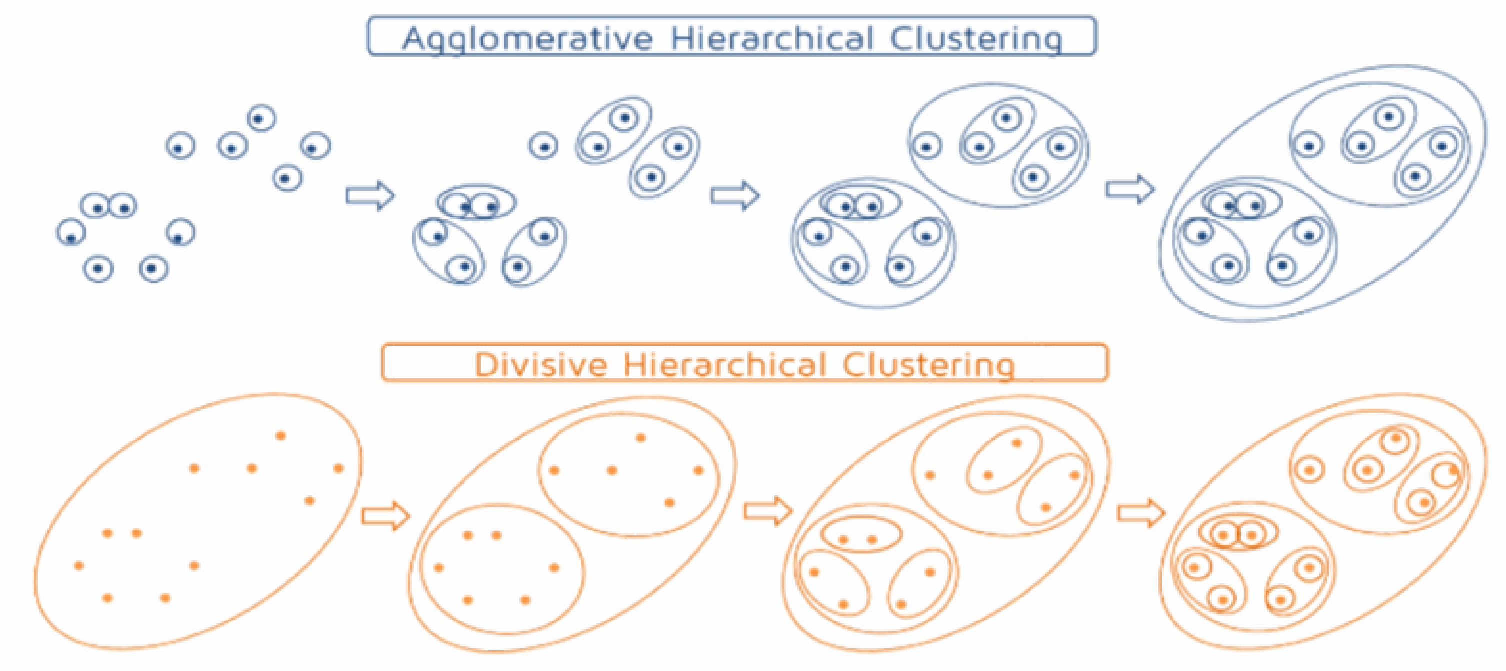
- May need to run several times to avoid suboptimal solutions
- Need to specify the number of clusters
- Not stable
  - Varying sizes, different densities, …
  - Even if we know the “right” number of clusters, k-means might not always recover them
- Boundary issues



Hierarchical Clustering

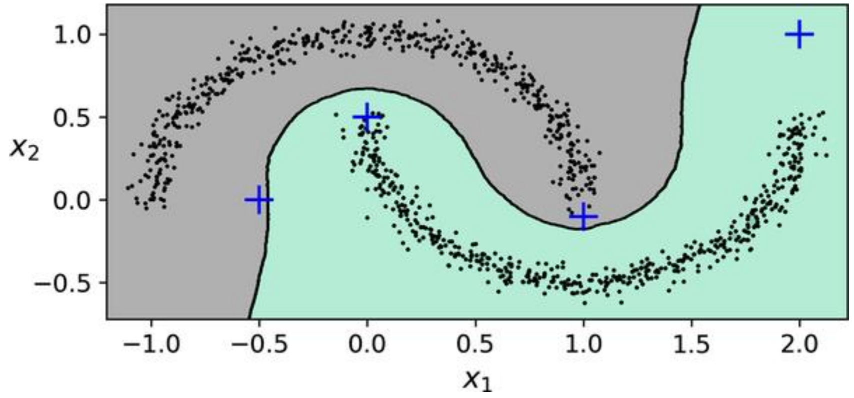
- Bottom-up: agglomerative
  - First merge similar instances, incrementally build larger clusters out of smaller clusters
- Top-down: divisive
  - Start with all data points in one cluster, split based on proximity

Note

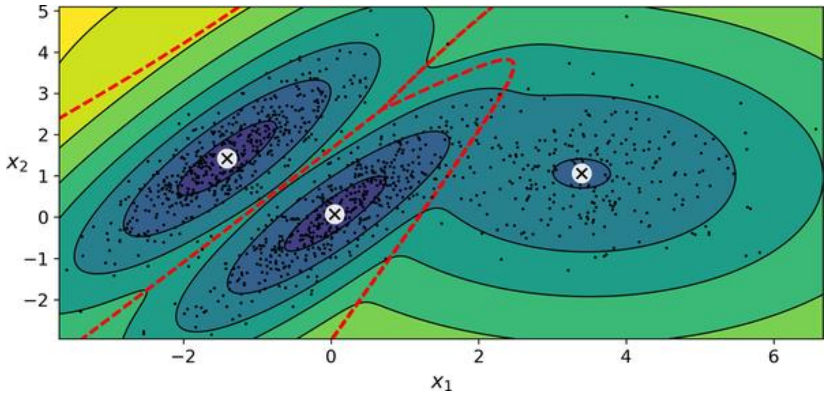


Other Clustering Techniques (Optional)

- DBSCAN: density-based spatial clustering of applications with noise
  - Define clusters as continuous regions of high density
  - Useful for arbitrary shapes



- GMM: Gaussian mixture model
  - Assume gaussian distribution for all instances
  - Useful for elliptical clusters



Lab 6

- Clustering of MNIST dataset
  - From Keras package
- Mini-batch version of KMeans
- External measure: true label

Competition

- **Pre-model thinking**: Why you chose the models and why they are appropriate for the problem
- **Model explanation**: Explain your data preprocessing and modeling approach
- **After-model interpretation**: Evaluate your model's performance

- **Evaluation Criteria**: model performance (30%) and presentation quality (70%)
  - How are you convinced by the presentation
- **Evaluation Link**: https://forms.gle/hcsn5F9SfdW2q3yY7

Next Week

- Reinforcement Learning
- Final Review

References

- Introduction to Machine Learning, Eric Xing and Ziv Bar-Joseph, School of Computer Science, Carnegie Mellon University
- Introduction to Machine Learning, David Sontag, New York University