Unsupervised Learning – Part 1

ESM3081 Programming for Data Science

Seokho Kang

성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)

Unsupervised Learning

Unsupervised Learning

- **Unsupervised Learning**
  - *(in general)* **Unlabeled training dataset** $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$, where each data point $\boldsymbol{x}_i = (x_{i1}, \dots, x_{id})$ contains *d* feature values
  - To find useful properties/patterns of the structures of the dataset

Unsupervised Learning

- ***Unlabeled Dataset***

Column: variable, attribute, **feature**, …

Input

Label

Row:
data point, **instance**, example, record, pattern, object, …

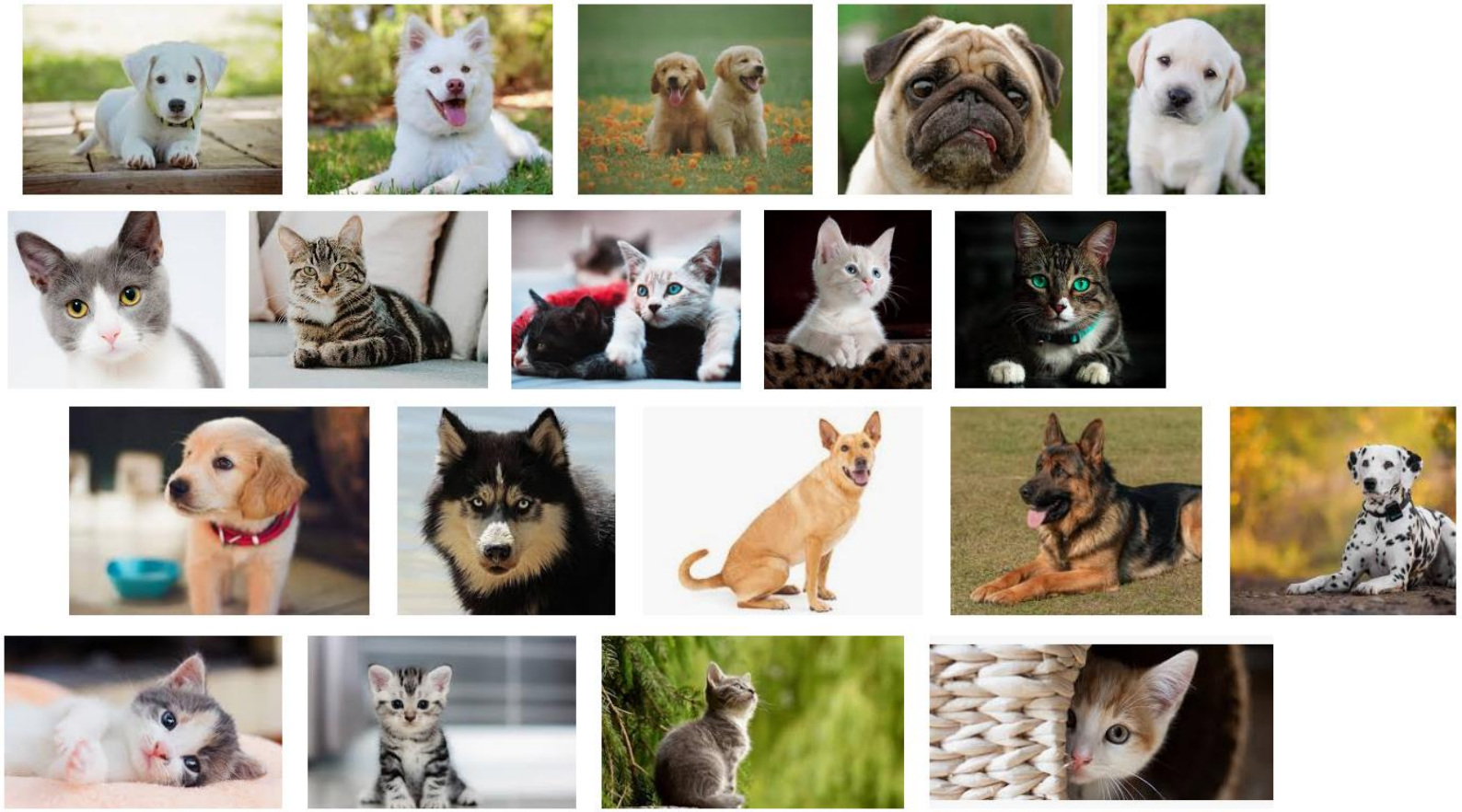
id	X_1	X_2	X_3	…	X_d
1	x_{11}	x_{12}	x_{13}	…	x_{1d}
2	x_{21}	x_{22}	x_{23}	…	x_{2d}
3	x_{31}	x_{32}	x_{33}	…	x_{3d}
4	x_{41}	x_{42}	x_{43}	…	x_{4d}
5	x_{51}	x_{52}	x_{53}	…	x_{5d}
6	x_{61}	x_{62}	x_{63}	…	x_{6d}
7	x_{71}	x_{72}	x_{73}	…	x_{7d}
…	…	…	…	…	…

X

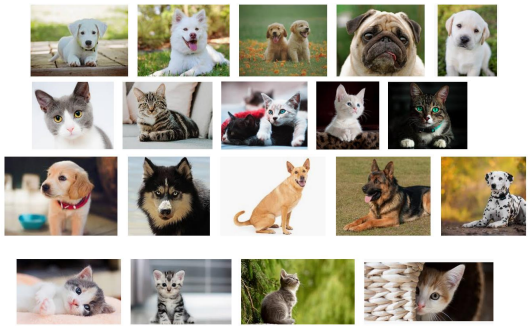
Unsupervised Learning

- Only the inputs are known, and no known outputs are given
- The unsupervised learning algorithm is just shown the input data and asked to extract knowledge from this data
- Unsupervised learning algorithms are usually harder to understand and evaluate

Types of Unsupervised Learning

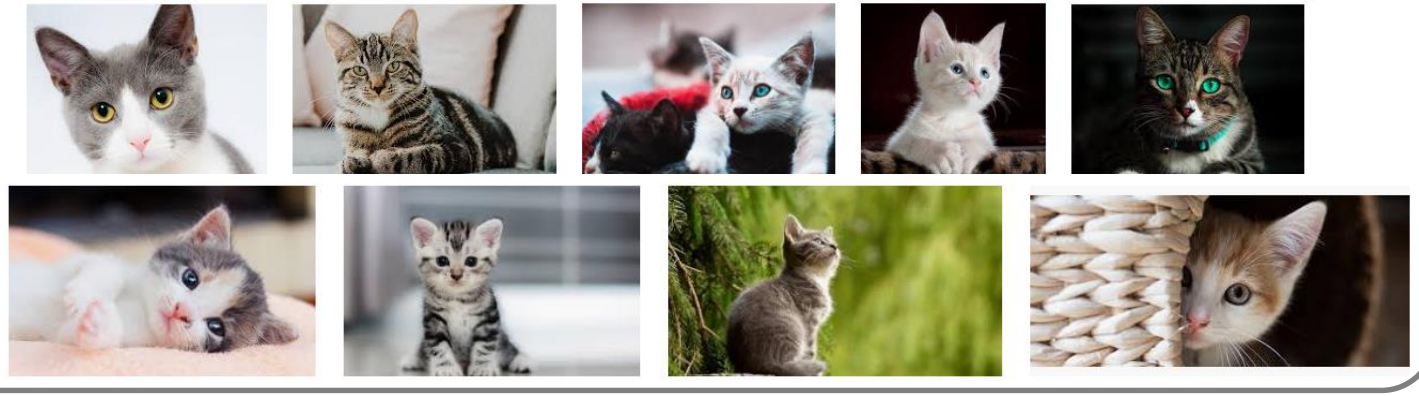


Types of Unsupervised Learning



Cluster 1

Cluster 2



Types of Unsupervised Learning

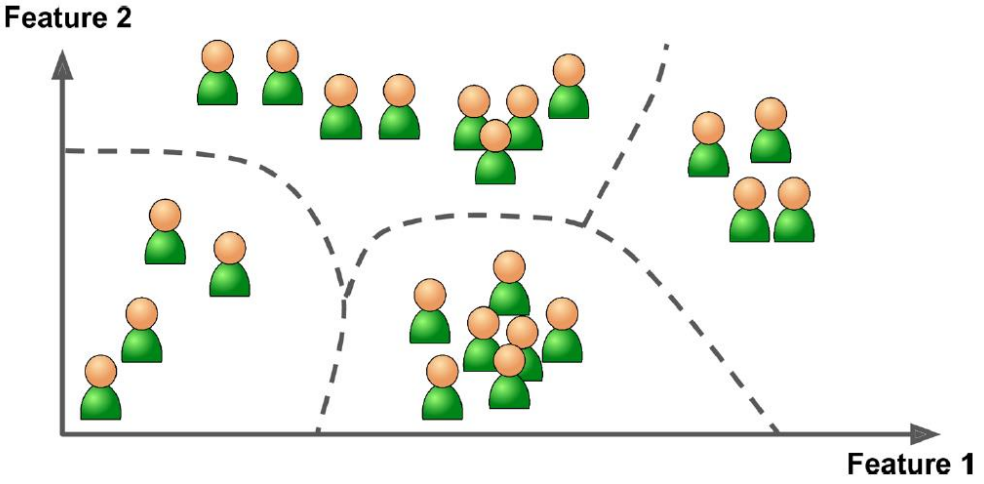
Types of Unsupervised Learning

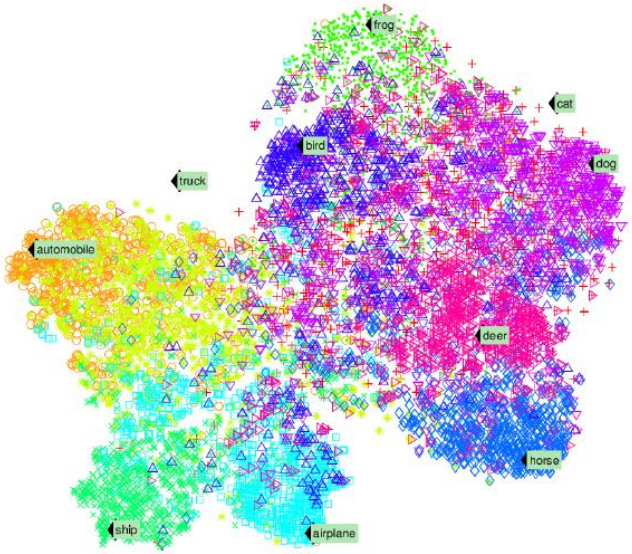
???

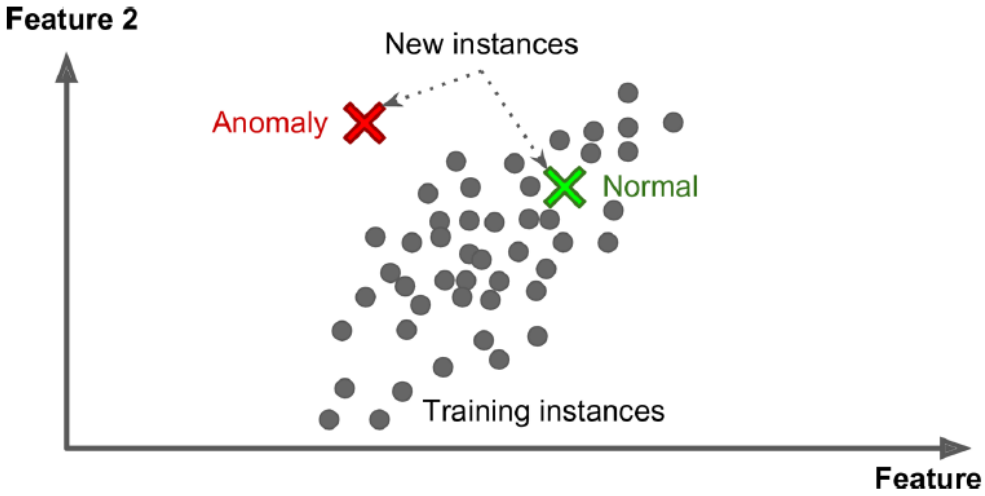
Types of Unsupervised Learning

- **Dimensionality Reduction**
  - Find a new way to represent this data that summarizes the essential characteristics with fewer features.
  - *Dimensionality reduction for visualization*: reduce to two or three dimensions for visualization purposes
- **Clustering**
  - Partition data into distinct groups of similar data points.
- **Anomaly Detection (One-Class Classification), Association Analysis, …**

Types of Unsupervised Learning







Challenges in Unsupervised Learning

- A major challenge in unsupervised learning is evaluating whether the algorithm learned something useful.
  - We don’t know what the right output should be.
  - It is very hard to tune the hyperparameters of an unsupervised learning algorithm.
  - The only way to evaluate the result is to inspect it manually.

- Unsupervised algorithms are used often in an exploratory setting
  - When a data scientist wants to understand the data better.
  - Rather than as part of a larger automatic system.

Learning algorithms covered in this course

- **Unsupervised Learning**
  - **Dimensionality Reduction & Visualization**
    - (Projection) Principal Component Analysis (PCA)
    - (Manifold Learning) t-distributed Stochastic Neighbor Embedding (t-SNE)
    - …
  - **Clustering**
    - K-Means
    - Hierarchical Clustering
    - DBSCAN
    - …