Unsupervised learning

Supervised Learning

- Data: labeled data, feature-response pairs (x, y)
- Goal: to learn a mapping from x to y
- Problem type:
  - Classification
  - Regression

Unsupervised Learning

- Data: *unlabeled* data, features x
- Goal: to uncover hidden structure, patterns, clusters, ...
- Typical tasks
  - **Clustering analysis:** cluster features or observations into groups with common traits
  - **Dimension reduction:** transform data from high-dimensional space to a low-dimensional space while preserving as much information as possible
  - ......

Reinforcement Learning

- Data: a sequence of state-action pairs and a cumulative reward
- Goal: to learn a policy (a mapping from state to action) that maximizes the cumulative reward.
- Typical applications:
  - Game-playing (e.g., AlphaGo, Dota2),
  - Robotics control
  - ......

Quick Summary

Aspect	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Goal	Learn a mapping from x to y	Discover hidden structure/patterns in x	Learn a policy that maximizes reward

Aspect	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Data	Labeled data	Unlabeled data	State-action pairs, and delayed rewards
Typical output	Classification or regression model	Clusters, low-dimensional embeddings	Policy

Clustering Analysis

- Popular techniques in analysis of complex data (e.g., genomics data)
- Aims to detecting homogeneous subgroups among the observations or the features
- Identities of the learned subgroups are unknown, but can often lead to meaningful interpretation after linked with additional information.
- We can cluster observations or features.
  - observations: clustering cancer samples to find similar cancer sub-types
  - features: clustering genes based on similar functions

Meaningful clusters

Without the groundtruth labels, how do we claim that our clusters results are meaningful?

Link to a clustering playground

Similarity and dissimilarity measures

Often we have to come up with numeric metrics to measure the similarity or dissimilarity between observations

- Similarity measures
  - Pearson correlation
  - Spearman correlation
  - Kendall's τ
- Dissimilarity measures
  - Euclidean distance
  - Manhattan distance

Similarity measures

When the features are numeric:

- Pearson correlation: This is the "usual" correlation coefficient, and measures the **linear** association between X_i and X_j.

- Spearman correlation. This is essentially the Pearson correlation applied to **ranked** observations.
- Kendall's τ. This correlation coefficient uses directly rankings among pairs of observations.
- ......

When the features are not numeric...

Dissimilarity measure

- Euclidean distance: sum of squared difference between feature values (ℓ_2-norm)

$$d_E(i,j) = \left[\sum_{l=1}^{p}(X_{i,l} - X_{j,l})^2\right]^{1/2}$$

- Manhattan distance: sum of absolute difference between feature values (ℓ_1-norm)

$$d_M(i,j) = \sum_{l=1}^{p}|X_{i,l} - X_{j,l}|$$

- ......

Checkout this example.

Common Clustering Methods

- Hierarchical clustering
  - Sequence of solutions organized in a hierarchical tree structure, called the dendrogram
  - Maintain a hierarchical structure

  When the number of clusters decrease, observations in the same cluster stays in the same cluster
- k-Means clustering
  - One of the most popular clustering method
  - Computationally efficient
  - No hierarchy among clusters

    When k decreases, observations in the same clusters might be reassigned to different clusters
- And many other methods

Dimension Reduction

- Goal: to transform data from a high-dimensional space to a low-dimensional space, without losing too much information
- Versatile tools for
  - Exploratory data analysis (e.g., tSNE)
  - Data processing for downstream analysis (e.g., principal component regression)
  - Uncover true latent pattern (e.g., intrinsic dimension of neural activities)
- Typical methods:
  - Principal Component Analysis (PCA)
  - t-distributed Stochastic Neighbor Embedding (tSNE)
