CS 307
Spring 2024
Dalpiaz

Unsupervised Learning
Supervised Learning

\[ y = f(X) \]

\[ \text{Features} \]

\[ \text{Target} \]

Supervision

Predict what class here?
Unsupervised Learning

- Dimension Reduction: PCA
- Clustering: K-means, Hierarchical Clustering, DBSCAN
- Density Estimation: KDE, Mixtures
- Outlier Detection: One-Class SVM, Isolation Forest
Dimensional Reduction

\[
\begin{bmatrix}
X_1 & X_2 & \ldots & X_p \\
\vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
X_1^* & \ldots & X_p^* \\
\vdots & \ddots & \vdots \\
\vdots & \ddots & \vdots \\
\end{bmatrix}
\]

Often used as preprocessing for supervised learning. **Example: PCA**

Can be used in a pipeline with `sklearn`.
Clustering

\[ x_1 \quad x_2 \]
\[ \vdots \quad \vdots \]
\[ \vdots \quad \vdots \]
\[ \vdots \quad \vdots \]
\[ \vdots \quad \vdots \]

Cluster Assignment

\[ \begin{array}{c}
A \\
B \\
B \\
C \\
B \\
A \\
C \\
\end{array} \]

Examples

- k-means
- hierarchical

- No standard to compare to in practice
- Order/labels meaningless
Clustering

Input

Output

Colors have no meaning!
Could be swapped!

Only groupings matter
Clustering

- How many clusters?

Often needs to be set by user before fitting.
Density Estimation

\[ f(x) \]

Examples

KDE

Mixtures Models

Could be used to generate new data
Outlier Detection
Outlier Detection

1. When should outliers be removed from training data? Only very carefully!

2. Eliminate "outliers" at test time?
   - Novelty detector
   - Useful within supervised pipelines
Novelty Detection

 Allow Predictions Here

 Refuse to Predict Here