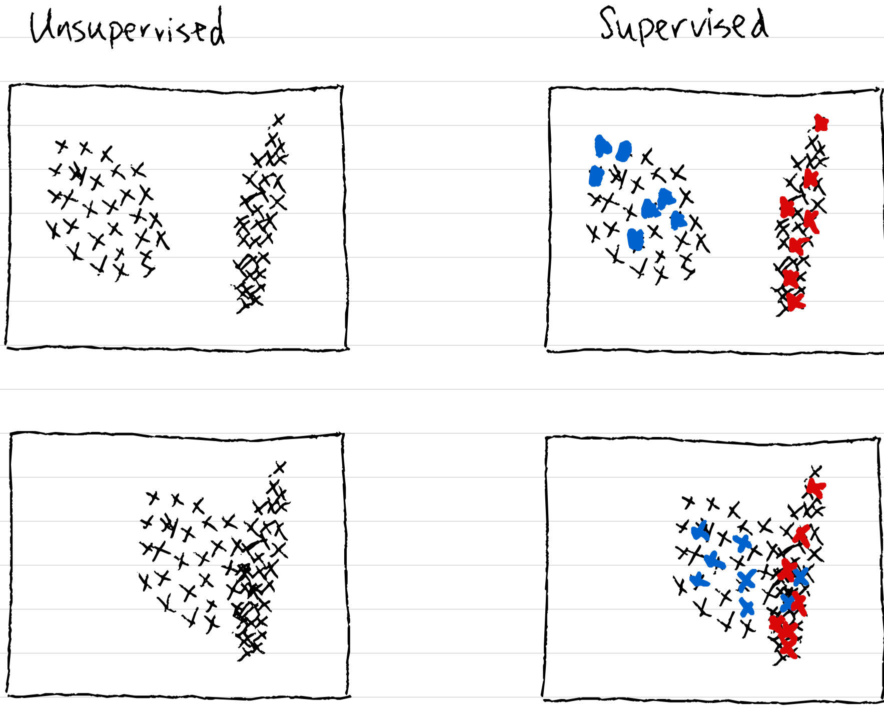
ME 491 Lecture 17

5.2 Supervised vs. Unsupervised Learning



* Unsupervised Learning

Input:

data $\{x_j \in \mathbb{R}^n, j \in Z := \{1, 2, \dots, m\}\}$

Output:

labels $\{y_j \in \{\pm 1\}, j \in Z\}$

Focused on providing labels y_j for all data. Generally, we use a subset of data D' to generate labels, and apply to data D more broadly.

* Supervised Learning

Input:

data $\{x_j \in \mathbb{R}^n, j \in Z := \{1, 2, 3, \dots, m\}\}$

labels $\{y_j \in \{\pm 1\}, j \in Z' \subset Z\}$

Output:

labels $\{y_j \in \{\pm 1\}, j \in Z\}$

Using the examples from Ch.5.1, we can formulate two classification problem.

- Fisher iris data

  $x_j = \{$sepal length, sepal width, petal length, petal width$\}$

  $y_j = \{$setosa, versicolor, virginica$\}$

  $D' \in \{$150 iris samples: 50 setosa, 50 versicolor, virginica$\}$

  $D \in \{$all setosa, versicolor, virginica irises in the world$\}$

- Dog cat data

  $x_j = \{64 \times 64$ image $= 4096$ pixels$\}$

  $y_j = \{$dog, cat$\} = \{1, -1\}$

  $D' = \{$160 image samples: 80 dogs and 80 cats$\}$

  $D = \{$all dogs and cats in the world$\}$

5.3 Unsupervised Learning: K-means clustering

K-means Clustering algorithm: one of the most prominent unsupervised algorithm.

Goal: partition m observations into K clusters. Each observation is labeled as belonging to a cluster with the nearest mean.

Protocol for K-means:

1. given initial values of K distinct means, compute the distance of each observation x_j to each of the K-means
2. label each observation as to the closest mean
3. After labeling, find the center-of-mass (mean) for each group (cluster)
4. repeat step 1-3 till convergence.

We can formulate this protocol into an optimization problem.

$$\underset{\mu_j}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{x_n \in D_j'} \| x_n - \mu_j \|^2$$

μ_j: mean of the jth cluster, D_j': subdomain of data of cluster j.

Graphical illustration of K-means