Unsupervised Learning

- So far discussed supervised learning:
  - Examples $(\mathbf{x}, y)$ are input-target pairs in $\mathcal{X} \times \mathcal{Y}$
  - Learning amounts to learning a mapping $h : \mathcal{X} \to \mathcal{Y}$
  - Loss measures discrepancy between $y$ and $\hat{y} = h(\mathbf{x})$, $\ell(y, \hat{y})$

- Sometimes we have plentiful of instances $\mathbf{x}_i$

  $$ S = \left\{ (\mathbf{x}_i, \cdot) \right\}_{i=1}^{m} \cup \left\{ (\mathbf{x}_i, y_i) \right\}_{i=1}^{n} $$

  ... but only handful of labels $m \gg n$
  ... or none at all $n = 0$

- It is nonetheless useful to find “structure” or meaningful patterns in the data
fMRI Data
Goals of Clustering

- Intuitively, grouping a set of objects (instances) such that
  - similar instances end up in the same cluster
  - dissimilar instances into different groups

- Imprecise & potentially ambiguous definition

- Disappointingly, not at all simple to define rigorously
Sources of Difficulty

• Inherit problem: lack of “ground truth” and tangible objective

• Technical difficulty:
  • Similarity & distance functions are not transitive

\[ \|u - v\| \leq \varepsilon \land \|v - w\| \leq \varepsilon \not\Rightarrow \|u - w\| \leq \varepsilon \]

• Cluster membership is transitive
  • Define \( u \sim v \) iff \( u \) and \( v \) belong to the same cluster
  • Then, \( u \sim v \land v \sim w \Rightarrow u \sim w \)
Clustering is Ambiguous

similar objects in same cluster  dissimilar objects are separated
Lack of Ground Truth

Partition points into two clusters:

We have two well justifiable solutions:
Model

- Input: set of elements $S = \{x_i\}_{i=1}^m$ where $x_i \in \mathcal{X} \subset \mathbb{R}^d$

- Distance $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ or similarity $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $s$ might not be symmetric, $s(u, v) \neq s(v, u)$

- Output: partition $\mathcal{C} = \{C_i\}_{i=1}^k$ of training set $S$ such that

$$S = \bigcup_{i=1}^k C_i \quad \text{and} \quad C_i \cap C_j = \emptyset$$

- Target number of clusters $k$ may be part of input or unknown
Cost-based Clustering

- Focus on distance-based \(d(u, v)\) clustering
- NP-hard problems, methods are prone to local minima
- Cost of partitioning \(\mathcal{C} = \{C_i\}_{i=1}^k\) of \(S\)?
- Define indicator
  \[
  1[i, j|C] = \begin{cases} 
  +1 & \exists r : x_i \in C_r \land x_j \in C_r \\
  -1 & \text{o.w.}
  \end{cases}
  \]
- Penalize for large intra-cluster & small inter-cluster distances
  \[
  \ell(S, \mathcal{C}) = \sum_{i,j=1}^{\left| S \right|} 1[i, j|C] \cdot d(x_i, x_j)
  \]
- Number of instances to compare \(O(n^2)\)

\[
|C_i| \approx \frac{n}{k} \Rightarrow \binom{k}{2} \left(\frac{n}{k}\right)^2 \equiv O(n^2) \quad \text{inter-cluster pairs}
\]
\[
k\left(\frac{n}{k}\right) \equiv O\left(\frac{n^2}{k}\right) \quad \text{intra-cluster pairs}
\]
k-Center Clustering

- Centroid-based clustering: intuitive, transitive, “aesthetic”

- Associate a center \( w_j \in \mathbb{R}^d \) with partition \( C_j \)

\[
x_i \in C_j \iff \forall l \neq j : d(x_i, w_j) < d(x_i, w_l)
\]

- Induces partition

\[
C_j = \{ i : \forall l \neq j d(x_i, w_j) < d(x_i, w_l) \}
\]

- Loss of k-centers

\[
\ell(S, C) = \sum_{j=1}^{k} \sum_{i \in C_j} d(x_i, w_j) = \sum_{i=1}^{m} \min_{j=1}^{k} d(x_i, w_j)
\]
Example of 3-Center Clustering
Skeleton of Metric Clustering

- Intialize each $w_j^0$ to a vector in $\mathbb{R}^d$

- For $t = 1, \ldots, T$
  - Associate each $x_i$ with its nearest centroid
    \[ \forall i : a^t(i) = \arg\min_{j=1}^{k} d(x_i, w_j^{t-1}) \]
  - Restimate centroids from associations
    \[ \forall j : w_j^t = \min_{w} \sum_{i : a^t(i) = j} d(x_i, w) \]
  - If $\forall i : a^t(i) = a^{t-1}(i)$ break
Convergence of Metric-based Clustering

- Centers at iteration $t$

  $$\mathcal{W}^t = \{w_j^t\}_{j=1}^k$$

- Partition at iteration $t$

  $$\mathcal{A}^t = \{a^t(i)\}_{i=1}^m$$

- Loss of partition and centers

  $$\ell(S, \mathcal{A}, \mathcal{W}) = \frac{1}{m} \sum_{i=1}^m d(x_i, w_{a(i)})$$

- Then, $\ell(S, \mathcal{A}^{t-1}, \mathcal{W}^{t-1}) > \ell(S, \mathcal{A}^t, \mathcal{W}^{t-1}) > \ell(S, \mathcal{A}^t, \mathcal{W}^t)$

- Since $\ell(S, \mathcal{A}, \mathcal{W}) \geq 0$ and $\forall t, j : w_j^t \in \bar{S}$

  $$\Rightarrow \ell(S, \mathcal{A}^t, \mathcal{W}^t)$$ converges to a local minimum
k-Means

- Use \( d(u, v) \overset{\text{def}}{=} \|u - v\|^2 \)

- Solving \( \min_w \sum_{i: a(i)=j} \|x_i - w\|^2 \) amounts to

  \[
  w_j = \frac{1}{n_j} \sum_{i: a(i)=j} x_i \quad \text{where} \quad n_j \overset{\text{def}}{=} |\{i : a(i) = j\}|
  \]

- Namely, center of mass of examples in cluster

- Runtime is: \( T \propto n \)
k-Medians

- Use $d(u, v) \overset{\text{def}}{=} \|u - v\|_1$

- Solving $\min_w \sum_{i:a(i)=j} \|x_i - w\|_1$ amounts to

  \[ w_j[r] = \min_\omega \sum_{i:a(i)=j} |x_i[r] - \omega| \]

  \[ = \text{median}\{x_i[r] : a(i) = j\} \]

- $w_j[r]$ is median of $r$’th coordinate of examples in cluster

- Runtime is: $T_{kn}$
Tricks & Treats

• Initialization:
  • At random
  • Agglomeratively: warm-start from $k - 1$ clusters
  • Agglomeratively: hierarchical from $2 \times \frac{k}{2}$ clusters
  • Using other clustering methods (e.g. spectral)

• Art of choosing number of clusters $k$ ...

• Small amounts of labeled data:
  • Determine number of clusters
  • Good initialization
  • Metric adjustment prior to clustering
Data Generated by $k$ Gaussians
Clustering with $\hat{k} = 3$
Why are the decision boundaries straight?
Clustering with $\hat{k} = 4$
Clustering with $\hat{k} = 5$
Original Means of Clusters