COS324: Introduction to Machine Learning
Lecture 18: Clustering

Prof. Elad Hazan & Prof. Yoram Singer

December 13, 2017

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 $\hat{y} = h(\mathbf{x})$, $\ell(y, \hat{y})$

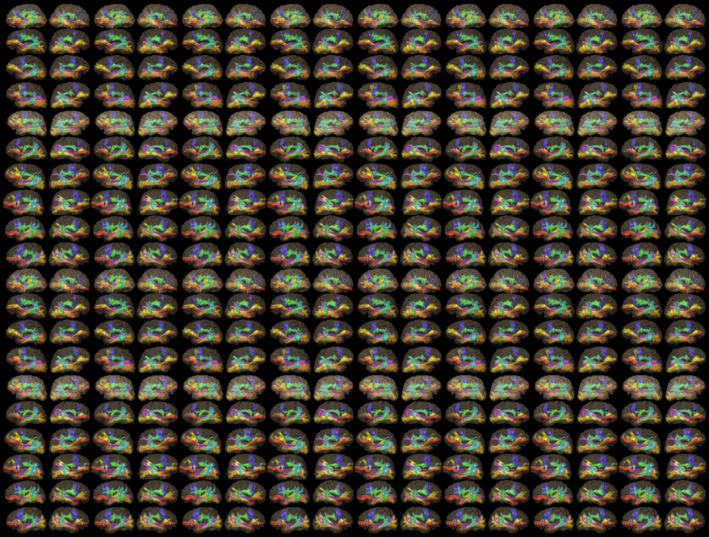
- Sometimes we have plentiful of instances $\mathbf{x}_i$

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

… 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

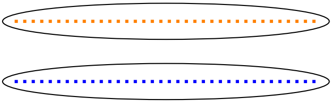
- Inherit problem: lack of "ground truth" and tangible objective
- Technical difficulty:
  - Similarity & distance functions are not transitive

$$\|\mathbf{u} - \mathbf{v}\| \leq \varepsilon \ \wedge \ \|\mathbf{v} - \mathbf{w}\| \leq \varepsilon \quad \not\Rightarrow \quad \|\mathbf{u} - \mathbf{w}\| \leq \varepsilon$$

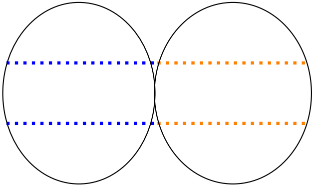
  - Cluster membership is transitive
    - Define $\mathbf{u} \sim \mathbf{v}$ iff $\mathbf{u}$ and $\mathbf{v}$ belong to the same cluster
    - Then, $\mathbf{u} \sim \mathbf{v} \ \wedge \ \mathbf{v} \sim \mathbf{w} \quad \Rightarrow \quad \mathbf{u} \sim \mathbf{w}$

Clustering is Ambigious

similar objects in same cluster

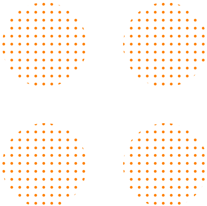


dissimilar objects are separated

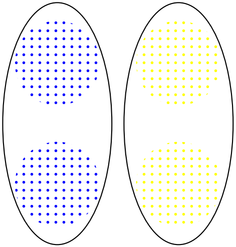


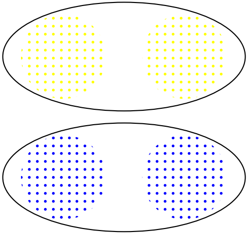
Lack of Ground Truth

Partition points into **two** clusters:



We have two well justifiable solutions:





Model

- Input: set of elements $S = \{x_i\}_{i=1}^m$ where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$
- 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, $s(\mathbf{u}, \mathbf{v}) \neq s(\mathbf{v}, \mathbf{u})$
- Output: partition $\mathcal{C} = \{C_i\}_{i=1}^k$ of training set S such that

$$S = \bigcup_{i=1}^{k} C_i \qquad\qquad C_i \cap C_j = \emptyset$$

- Target number of clusters k may be part of input or unkown

Cost-based Clustering

- Focus on distance-based $d(\mathbf{u}, \mathbf{v})$ clustering
- NP-hard problems, methods are prone to local minima
- Cost of partitioning $\mathcal{C} = \left\{C_i\right\}_{i=1}^{k}$ of S ?
- Define indicator

$$\mathbb{1}[i, j|\mathcal{C}] = \begin{cases} +1 & \exists r : \mathbf{x}_i \in C_r \wedge \mathbf{x}_j \in C_r \\ -1 & \text{o.w.} \end{cases}$$

- Penalize for large intra-cluster & small inter-cluster distances

$$\ell(S, \mathcal{C}) = \sum_{i,j=1}^{|S|} \mathbb{1}[i, j|\mathcal{C}]\, d(\mathbf{x}_i, \mathbf{x}_j)$$

- Number of instances to compare $O\big(n^2\big)$

$$|C_i| \approx \frac{n}{k} \;\Rightarrow\; \begin{array}{ll} \binom{k}{2}\left(\frac{n}{k}\right)^2 \;\equiv\; O\big(n^2\big) & \text{inter-cluster pairs} \\ k\binom{\frac{n}{k}}{2} \;\equiv\; O\Big(\frac{n^2}{k}\Big) & \text{intra-cluster pairs} \end{array}$$

k-Center Clustering

- Centroid-based clustering: intuitive, transitive, "aesthetic"
- Associate a center $\mathbf{w}_j \in \mathbb{R}^d$ with partition C_j

$$\mathbf{x}_i \in C_j \quad \Leftrightarrow \quad \forall l \neq j: \; d(\mathbf{x}_i, \mathbf{w}_j) < d(\mathbf{x}_i, \mathbf{w}_l)$$

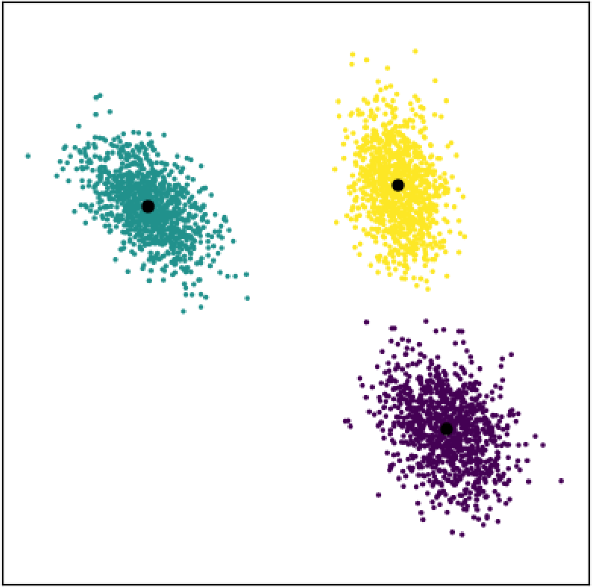
- Induces partition

$$C_j = \left\{ i : \forall l \neq j \; d(\mathbf{x}_i, \mathbf{w}_j) < d(\mathbf{x}_i, \mathbf{w}_l) \right\}$$

- Loss of k-centers

$$\ell(S, \mathcal{C}) = \sum_{j=1}^{k} \sum_{i \in C_j} d(\mathbf{x}_i, \mathbf{w}_j) \; = \; \sum_{i=1}^{m} \min_{j=1}^{k} d(\mathbf{x}_i, \mathbf{w}_j)$$

Example of 3-Center Clustering



Skeleton of Metric Clustering

- Intialize each $\mathbf{w}_j^0$ to a vector in $\mathbb{R}^d$
- For $t = 1, \ldots, T$
  - Associate each $\mathbf{x}_i$ with its nearest centroid

$$\forall i : \; a^{\mathrm{t}}(i) = \arg\min_{j=1}^{k} d(\mathbf{x}_i, \mathbf{w}_j^{\mathrm{t}-1})$$

  - Restimate centroids from associations

$$\forall j : \; \mathbf{w}_j^{\mathrm{t}} = \min_{\mathrm{w}} \sum_{i:a^{\mathrm{t}}(i)=j} d(\mathbf{x}_i, \mathbf{w})$$

  - If $\forall i : \; a^{\mathrm{t}}(i) = a^{\mathrm{t}-1}(i)$ break

Convergence of Metric-based Clustering

- Centers at iteration t

$$\mathcal{W}^t = \left\{\mathbf{w}_j^t\right\}_{j=1}^k$$

- Partition at iteration t

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

- Loss of partition and centers

$$\ell(S, \mathcal{A}, \mathcal{W}) = \frac{1}{m}\sum_{i=1}^m d(\mathbf{x}_i, \mathbf{w}_{a(i)})$$

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

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

  $\Rightarrow \quad \ell\big(S, \mathcal{A}^{\mathrm{t}}, \mathcal{W}^{\mathrm{t}}\big)$ converges to a local minimum

k-Means

- Use $d(\mathbf{u}, \mathbf{v}) \stackrel{\text{def}}{=} \|\mathbf{u} - \mathbf{v}\|_2^2$

- Solving $\min_{\mathrm{w}} \sum_{i:a(i)=j} \|\mathbf{x}_i - \mathbf{w}\|^2$ amounts to

$$\mathbf{w}_j = \frac{1}{n_j} \sum_{i:a(i)=j} \mathbf{x}_i \quad \text{where} \quad n_j \stackrel{\text{def}}{=} \left|\{i : a(i) = j\}\right|$$

- Namely, center of mass of examples in cluster

- Runtime is: T k n

k-Medians

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

- Solving $\min_{\mathrm{w}} \sum_{i:a(i)=j} \|\mathbf{x}_i - \mathbf{w}\|_1$ amounts to

$$
\begin{aligned}
\mathbf{w}_j[r] &= \min_{\omega} \sum_{i:a(i)=j} \left|\mathbf{x}_i[r] - \omega\right| \\
&= \mathbf{median}\left\{\mathbf{x}_i[r] \;:\; a(i) = j\right\}
\end{aligned}
$$

- $\mathbf{w}_j[r]$ is median of r'th coordinate of examples in cluster

- Runtime is: T k n

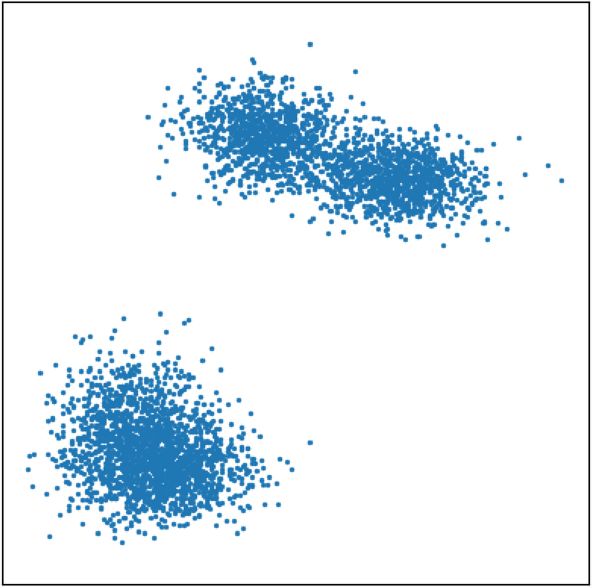
Tricks & Treats

- 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)

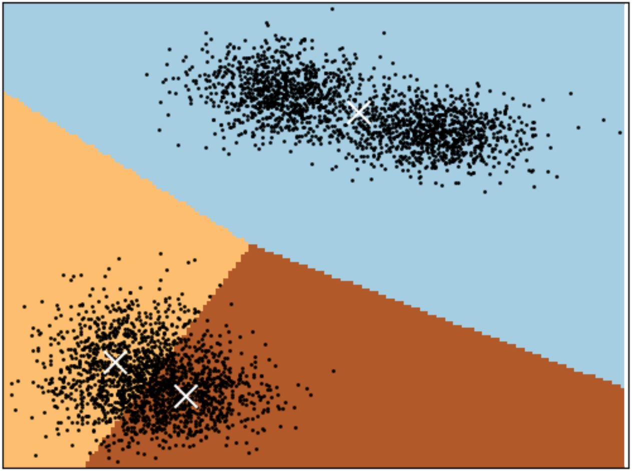
- Art of choosing number of clusters k ...

- Small amounts of labeled data:
    - Determine number of clusters
    - Good initialization
    - Metric adjustment prior to clustering

Data Generated by k Gaussians

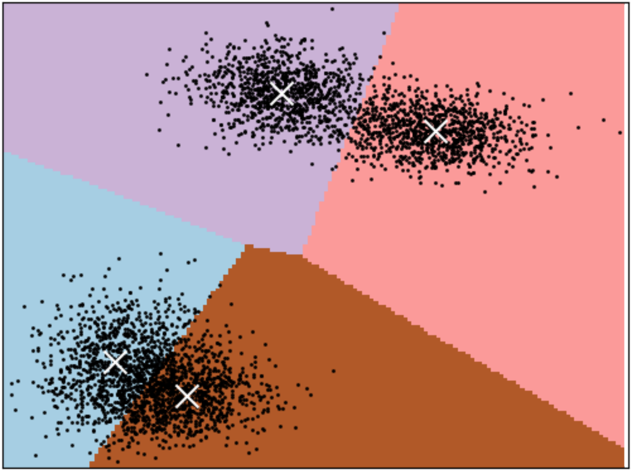


Clustering with $\hat{k} = 3$

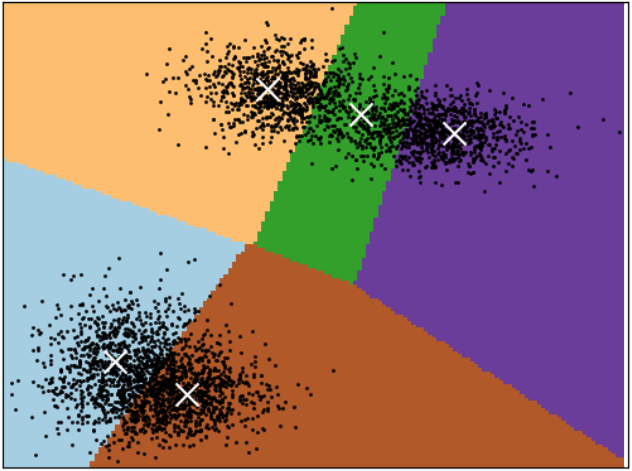


Why are the decision boundaries straight?

Clustering with $\hat{k} = 4$



Clustering with $\hat{k} = 5$



Original Means of Clusters