Unsupervised Learning

Ravi Kothari, Ph.D.
ravi.kothari@ashoka.edu.in

"If all you do is follow the herd, you'll just be stepping in poop all day - Wayne Dyer"

Unsupervised Learning

Often times, we have N (possibly noisy) observations $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$. The desired output corresponding to each pattern is not known - perhaps it is expensive to obtain them, perhaps we do not know the desired outputs i.e. the patterns are unlabeled. Is it possible to induce anything from such data?

Unsupervised Learning

Often times, we have N (possibly noisy) observations $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$. The desired output corresponding to each pattern is not known - perhaps it is expensive to obtain them, perhaps we do not know the desired outputs i.e. the patterns are unlabeled. Is it possible to induce anything from such data?

- We can partition the patterns into cohesive groups (clustering). Patterns in the same group are *similar* to each other and *dis-similar* from the patterns in other groups

Unsupervised Learning

Often times, we have N (possibly noisy) observations $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$. The desired output corresponding to each pattern is not known - perhaps it is expensive to obtain them, perhaps we do not know the desired outputs i.e. the patterns are unlabeled. Is it possible to induce anything from such data?

- We can partition the patterns into cohesive groups (clustering). Patterns in the same group are *similar* to each other and *dis-similar* from the patterns in other groups
  - (Dis-)Similarity implies there is a way of measuring the distance between $x^{(i)}$ and $x^{(j)}$

Unsupervised Learning

Unsupervised Learning

- In Euclidean space for example,

$$\| x^{(i)} - x^{(j)} \| = \sqrt{[x^{(i)} - x^{(j)}]^T [x^{(i)} - x^{(j)}]}$$

$$cos(\psi) = \frac{{x^{(i)}}^T x^{(j)}}{\| x^{(i)} \| \quad \| x^{(j)} \|}$$

Unsupervised Learning

- In Euclidean space for example,

$$\| x^{(i)} - x^{(j)} \| = \sqrt{[x^{(i)} - x^{(j)}]^T [x^{(i)} - x^{(j)}]}$$

$$cos(\psi) = \frac{{x^{(i)}}^T x^{(j)}}{\| x^{(i)} \| \quad \| x^{(j)} \|}$$

- In the case of text documents, one can convert the document in to vector. For example, a vector of length equal to the size of the vocabulary. Each position corresponds to a word, and is 1 if the word appears in the document. One can then use one of the above distance measures amongst the document vectors.

Unsupervised Learning

- In Euclidean space for example,

$$\| x^{(i)} - x^{(j)} \| = \sqrt{[x^{(i)} - x^{(j)}]^T [x^{(i)} - x^{(j)}]}$$

$$cos(\psi) = \frac{{x^{(i)}}^T x^{(j)}}{\| x^{(i)} \| \quad \| x^{(j)} \|}$$

- In the case of text documents, one can convert the document in to vector. For example, a vector of length equal to the size of the vocabulary. Each position corresponds to a word, and is 1 if the word appears in the document. One can then use one of the above distance measures amongst the document vectors.
  - This is rather coarse and treats each word as important as another (see for example, TF-IDF that gives a better idea of the importance of a word). There are many other approaches

k-means Clustering

Given $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$, we wish to find k cluster centers $\{\mu^{(i)}\}_{i=1}^{k}$

k-means Clustering

Given $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$, we wish to find k cluster centers $\{\mu^{(i)}\}_{i=1}^{k}$

1. Initialize k clusters $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(k)}$ to random values (or assign them to be equal to k randomly chosen examples (observations))

k-means Clustering

Given $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$, we wish to find k cluster centers $\{\mu^{(i)}\}_{i=1}^{k}$

1. Initialize k clusters $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(k)}$ to random values (or assign them to be equal to k randomly chosen examples (observations))
2. Assign each $x^{(i)}$ to the closest cluster center

k-means Clustering

Given $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$, we wish to find k cluster centers $\{\mu^{(i)}\}_{i=1}^{k}$

1. Initialize k clusters $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(k)}$ to random values (or assign them to be equal to k randomly chosen examples (observations))
2. Assign each $x^{(i)}$ to the closest cluster center
3. Compute the centroid of each cluster based on the examples (observations) assigned to that cluster, i.e.,

$$\mu^{(i)} = \frac{1}{N^{(i)}} \sum_{l \in \mu^{(i)}} x^{(l)}$$

k-means Clustering

Given $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$, we wish to find k cluster centers $\{\mu^{(i)}\}_{i=1}^{k}$

1. Initialize k clusters $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(k)}$ to random values (or assign them to be equal to k randomly chosen examples (observations))
2. Assign each $x^{(i)}$ to the closest cluster center
3. Compute the centroid of each cluster based on the examples (observations) assigned to that cluster, i.e.,

$$\mu^{(i)} = \frac{1}{N^{(i)}} \sum_{l \in \mu^{(i)}} x^{(l)}$$

4. Go to Step 2 until $\mu^{(i)}$'s stop changing i.e. cluster assignment do not change

Example

Example

1. Let $x^{(1)} = [1 \quad 1]^T$, $x^{(2)} = [2 \quad 2]^T$, $x^{(3)} = [4 \quad 4]^T$, $x^{(4)} = [5 \quad 5]^T$. Let $k = 2$.

Example

1. Let $x^{(1)} = [1 \quad 1]^T$, $x^{(2)} = [2 \quad 2]^T$, $x^{(3)} = [4 \quad 4]^T$, $x^{(4)} = [5 \quad 5]^T$.
   Let $k = 2$.
2. Let $\mu^{(1)} \leftarrow x^{(1)}$, $\mu^{(2)} \leftarrow x^{(2)}$

Example

1. Let $x^{(1)} = [1 \quad 1]^T$, $x^{(2)} = [2 \quad 2]^T$, $x^{(3)} = [4 \quad 4]^T$, $x^{(4)} = [5 \quad 5]^T$. Let $k = 2$.
2. Let $\mu^{(1)} \leftarrow x^{(1)}$, $\mu^{(2)} \leftarrow x^{(2)}$
3. Assign each $x^{(l)}$ to the closest cluster center. So, $x^{(1)} \rightarrow \mu^{(1)}$, $x^{(2)} \rightarrow \mu^{(2)}$, $x^{(3)} \rightarrow \mu^{(3)}$, and $x^{(4)} \rightarrow \mu^{(4)}$

Example

1. Let $x^{(1)} = [1 \quad 1]^T$, $x^{(2)} = [2 \quad 2]^T$, $x^{(3)} = [4 \quad 4]^T$, $x^{(4)} = [5 \quad 5]^T$. Let $k = 2$.
2. Let $\mu^{(1)} \leftarrow x^{(1)}$, $\mu^{(2)} \leftarrow x^{(2)}$
3. Assign each $x^{(l)}$ to the closest cluster center. So, $x^{(1)} \rightarrow \mu^{(1)}$, $x^{(2)} \rightarrow \mu^{(2)}$, $x^{(3)} \rightarrow \mu^{(3)}$, and $x^{(4)} \rightarrow \mu^{(4)}$
4. Recompute the cluster centers. $\mu^{(1)} = 1/2(x^{(1)} + x^{(2)}) = [1.5 \quad 1.5]^T$ and $\mu^{(2)} = 1/2(x^{(3)} + x^{(4)}) = [4.5 \quad 4.5]^T$

Example

1. Let $x^{(1)} = [1 \quad 1]^T$, $x^{(2)} = [2 \quad 2]^T$, $x^{(3)} = [4 \quad 4]^T$, $x^{(4)} = [5 \quad 5]^T$. Let $k = 2$.
2. Let $\mu^{(1)} \leftarrow x^{(1)}$, $\mu^{(2)} \leftarrow x^{(2)}$
3. Assign each $x^{(l)}$ to the closest cluster center. So, $x^{(1)} \rightarrow \mu^{(1)}$, $x^{(2)} \rightarrow \mu^{(2)}$, $x^{(3)} \rightarrow \mu^{(3)}$, and $x^{(4)} \rightarrow \mu^{(4)}$
4. Recompute the cluster centers. $\mu^{(1)} = 1/2(x^{(1)} + x^{(2)}) = [1.5 \quad 1.5]^T$ and $\mu^{(2)} = 1/2(x^{(3)} + x^{(4)}) = [4.5 \quad 4.5]^T$
5. Go to step 3

Difficulties with k-means

Difficulties with k-means

- Need to know k *a priori*

Difficulties with k-means

- Need to know k *a priori*
- Clusters need to be spherical

Difficulties with k-means

- Need to know k *a priori*
- Clusters need to be spherical
- k-means minimizes a cost function of the type,

$$J = \sum_{i=1}^{N}\sum_{j=1}^{k} I(\mu^{(j)} - x^{(i)}) \times \| (\mu^{(j)} - x^{(i)}) \| \tag{1}$$

where, $I(\mu^{(j)} - x^{(i)})$ is 1 if $(\mu^{(j)}$ is the closest cluster center to $x^{(i)}$

Density Based Clustering

Density Based Clustering

- We adopt the definition that a cluster is a set of densely connected points

Density Based Clustering

- We adopt the definition that a cluster is a set of densely connected points
- We look at a density based clustering algorithm, DBScan, that can detect arbitrary shaped clusters and achieves the clustering based on a single pass through the data

DBScan

DBScan

- We define an ϵ-neighborhood as,

$$\mathcal{N}_\epsilon(x) = \{x' | d(x, x') \leq \epsilon\}$$

DBScan

- We define an ϵ-neighborhood as,

$$\mathcal{N}_\epsilon(x) = \{x' | d(x, x') \leq \epsilon\}$$

- We say the ϵ-neighborhood is dense if it contains at least m points

DBScan

- We define an ϵ-neighborhood as,

$$\mathcal{N}_\epsilon(x) = \{x' | d(x, x') \leq \epsilon\}$$

- We say the ϵ-neighborhood is dense if it contains at least m points
- Given ϵ and m, we can categorize each data point into 3 categories,

DBScan

- We define an ϵ-neighborhood as,

$$\mathcal{N}_\epsilon(x) = \{x' | d(x, x') \leq \epsilon\}$$

- We say the ϵ-neighborhood is dense if it contains at least m points
- Given ϵ and m, we can categorize each data point into 3 categories,
  - A *core point* has more than m points in its neighborhood

DBScan

- We define an ϵ-neighborhood as,

$$\mathcal{N}_\epsilon(x) = \{x' | d(x, x') \leq \epsilon\}$$

- We say the ϵ-neighborhood is dense if it contains at least m points
- Given ϵ and m, we can categorize each data point into 3 categories,
  - A *core point* has more than m points in its neighborhood
  - A *border point* has less than m points in its neighborhood but is the neighborhood of a core point

DBScan

- We define an ϵ-neighborhood as,

$$\mathcal{N}_\epsilon(x) = \{x' | d(x, x') \leq \epsilon\}$$

- We say the ϵ-neighborhood is dense if it contains at least m points
- Given ϵ and m, we can categorize each data point into 3 categories,
  - A *core point* has more than m points in its neighborhood
  - A *border point* has less than m points in its neighborhood but is the neighborhood of a core point
  - A *noise point* is neither a core point nor a border point

DBScan (contd.)

DBScan (contd.)

- *Directly Density-Reachable*: A point x' is directly density-reachable from a point x if x is a core point and x' is in the ϵ-neighborhood of x

DBScan (contd.)

- *Directly Density-Reachable*: A point x' is directly density-reachable from a point x if x is a core point and x' is in the ϵ-neighborhood of x
- *Density-Reachable*: x is directly density-reachable from x'. x' is directly density-reachable from x''. Then, x is (indirectly) density-reachable from x''

The DBScan Algorithm

Initialize all the weights to small random values
for i = 1 to N
    if $x^{(i)}$ is not yet classified
        if $x^{(i)}$ is a core object
            Collect all data points density reachable from $x^{(i)}$ and assign them to a new cluster
        else
            Assign $x^{(i)}$ to noise
end for