07_Unsupervised_Learning

Unsupervised Learning

'Unlabeled' data
e.g. Clustering, Density Estimation, Anomaly Detection

Clustering

High Intra-cluster Similarity
Low Inter-cluster Similarity

- **K-means Algorithm**

Given data set $D = \{x_1, x_2, \dots, x_n\}$, desired cluster $C = \{C_1, C_2, \dots, C_k\}$

$$E = \sum_{i=1^k} \sum_{\vec{x} \in C_i} \|\vec{x} - \vec{\mu}\|_2^2$$

$$\vec{\mu} = \frac{1}{|C_i|} \sum_{x \in C_i} \vec{x}$$

$\vec{\mu}$ is the mean vector of cluster C, the equation above indicates the closeness between samples in cluster and $\vec{\mu}$.

- **Algorithm**

input: $D = \{x_1, x_2, \dots, x_n\}$,number of desired cluster k
Procedure:
initialization: assign k initial values from D to k centroids
repeat $(Iteration)$
assign each point x_j to closest centroid c_j ;
$(Euclidean\ Distance: isotropic, favours\ spherical\ clusters.)$
$d(x, y) = \sqrt{\sum_1^n (x_n - y_n)^2}$

$for\ \ j = 1, 2, \dots, m\ \ do$
$\quad calculate\ distance\ between\ x_j\ and\ \vec{\mu}: d_{ij} = \|\vec{x}_j - \vec{\mu}_i\|_2$
$\quad the\ cloest\ d_{ij}\ to\ fix\ subscript\ of\ C: \lambda_j = argmin_{i \in \{1,2,\dots,k\}}\ d_{ji}$
$\quad assign\ sample\ x_j\ to\ corresponding\ C_{\lambda_j}: C_{\lambda_j} \cup \{x_j\}$
$end\ for$

#compute new centroids as mean of each group of points;
$for\ \ i = 1, 2, \dots, k\ \ do$
$\quad calculate\ new\ mean\ vector: \vec{\mu}_i' = \frac{1}{|C_i|} \sum_{x \in C_i} \vec{x}$
$\quad if\ \vec{\mu}_i' \neq \vec{\mu}_i\ then,\ \vec{\mu}_i = \vec{\mu}_i'$
$\quad else\ return\ \vec{\mu}_i$
$end\ for$

until
#centroids do not change;$(Converge)$
return k clusters;

output: cluster $C = \{C_1, C_2, \ldots, C_k\}$