Unsupervised Learning: Clustering

EE219: Large Scale Data Mining

Professor Roychowdhury

K-means clustering algorithm

Algorithm

0. Randomly initialize K cluster centers (centroids)
1. Iterate until convergence
   1.1 For each data point, find closest cluster center (partitioning step)
   1.2 Replace each centroid by average of data points in its partition

Objective function

Write $x_i = (x_{i1}, \ldots, x_{ip})$:

Let the centroids be denoted by $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_K$, $\bar{x}_i = \frac{1}{p}\sum_{j=1}^{p} x_{ij}$, and the clusters by $c_1, c_2, \ldots, c_K$, then the objective function of K-means is to minimize Euclidean distance of the points with the centroids of corresponding clusters (within cluster sum of squares):

$$\sum_{k=1}^{K} \sum_{i \in c_k} \|x_i - \bar{x}_k\|^2$$

- Consider the assignment function $C(i)$:

$$C : 1, 2, \ldots, N \rightarrow (1, 2, \ldots, K)$$

- K-means minimizes $W(C)$

$$\begin{aligned} W(C) &= \frac{1}{2} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{C(i)=k} \sum_{C(j)=k} \|x_i - x_j\|^2 \\ &= \sum_{k=1}^{K} \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2 \end{aligned}$$

A proof for this equivalence is given in the following slide

- K-means solves the following problem to find assignment function C:

$$\min_{C, \bar{x}_1 \ldots \bar{x}_k} \sum_{k} \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2$$

The outer summation ($k = 1$ through K) is over different clusters. The summands for each k are data within the cluster k. So we just prove the equivalence for each cluster k. In other words, we show that if the number of data points in cluster k is N_k, then:

$$\sum_{C(i)=k} \sum_{C(j)=k} \|x_i - x_j\|^2 = \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} \|x_i - x_j\| = 2N_k \sum_{i=1}^{N_k} \|x_i - \bar{x}\|^2$$

$$\begin{aligned}
&\sum_{i=1}^{N_k} \left(\sum_{j=1}^{N_k} \|x_i - x_j\|^2 \right) \\
&= \sum_{i=1}^{N_k} \left(\sum_{j=1}^{N_k} \|(x_i - \bar{x}) - (x_j - \bar{x})\|^2 \right) \\
&= \sum_{i=1}^{N_k} \left(\sum_{j=1}^{N_k} [(x_i - \bar{x}) - (x_j - \bar{x})]^T [(x_i - \bar{x}) - (x_j - \bar{x})] \right) \\
&= \sum_{i=1}^{N_k} \left(\sum_{j=1}^{N_k} \left(\|x_i - \bar{x}\|^2 + \|x_j - \bar{x}\|^2 - 2(x_i - \bar{x})^T (x_j - \bar{x}) \right) \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N_k} \left(\sum_{j=1}^{N_k} \left(\|x_i - \overline{x}\|^2 + \|x_j - \overline{x}\|^2 - 2(x_i - \overline{x})^T (x_j - \overline{x}) \right) \right) \\
&= \sum_{i=1}^{N_k} \left(N_k \|x_i - \overline{x}\|^2 + \sum_{j=1}^{n} \|x_j - \overline{x}\|^2 - 2 \sum_{j=1}^{n} (x_i - \overline{x})^T (x_j - \overline{x}) \right) \\
&= \sum_{i=1}^{N_k} \left(2N_k \|x_i - \overline{x}\|^2 - 2(x_i - \overline{x})^T \sum_{j=1}^{n} (x_j - \overline{x}) \right) \\
&= 2N_k \sum_{i=1}^{N_k} \left(\|x_i - \overline{x}\|^2 \right)
\end{aligned}
$$

Thus:

$$
\sum_{i=1}^{N_k} \left(\sum_{j=1}^{N_k} \|x_i - x_j\|^2 \right) = 2N_k \sum_{i=1}^{N_k} \|x_i - \overline{x}\|^2 \Rightarrow
$$

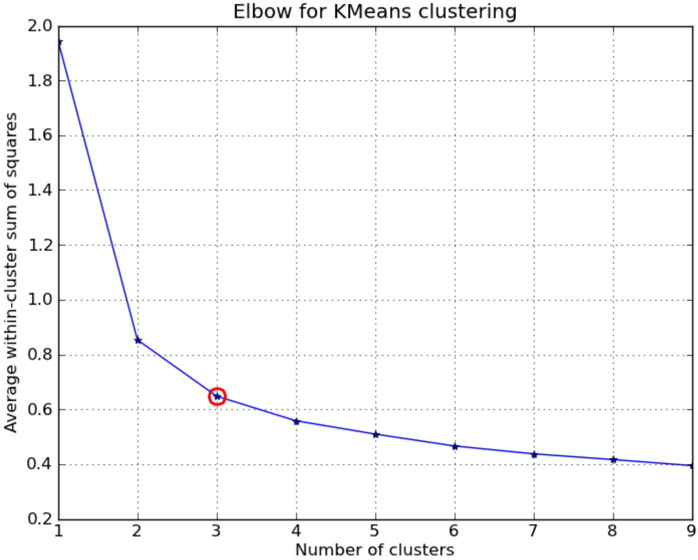
$$
\sum_{\substack{i,j=1 \\ i<j}}^{N_k} \|x_i - x_j\|^2 = N_k \sum_{i=1}^{N_k} \|x_i - \overline{x}\|^2
$$

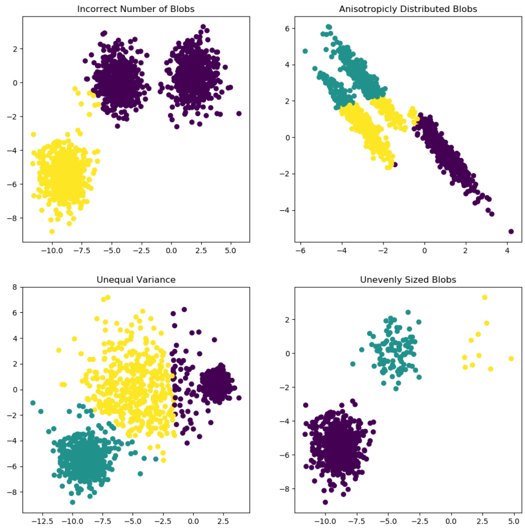
Initial centroid problem

- K-means converges to a local optimum whose quality largely depends on the initial choice of centroids
- Solution: multiple (e.g. 100) runs with random initial cluster centroids, then choosing the ones with the minimal final cost function

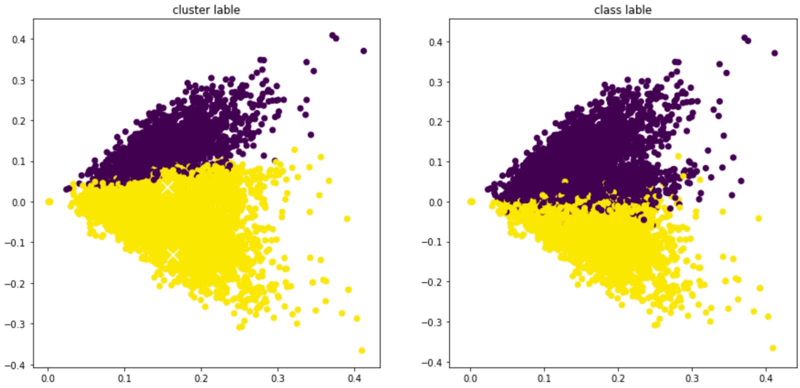
How to choose K?

- Elbow method





credit: http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html#sphx-glr-auto-examples-cluster-plot-kmeans-assumptions-py

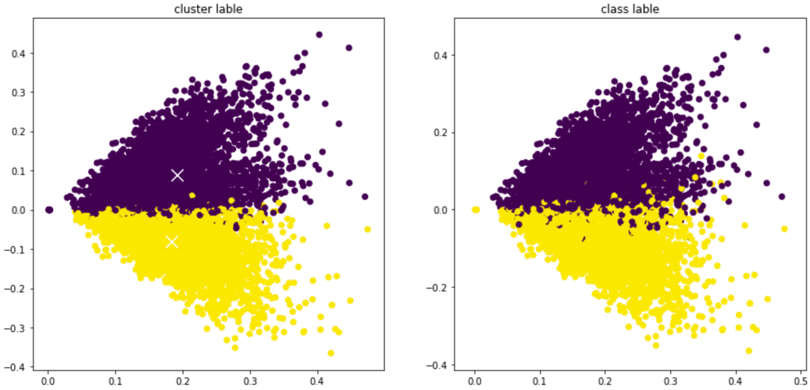


A non-ideal clustering result

Left: clustering results; Right: ground truth

Original data is high-dimensional; points are colored according to their labels

SVD is performed to project the data to 2D for visualization

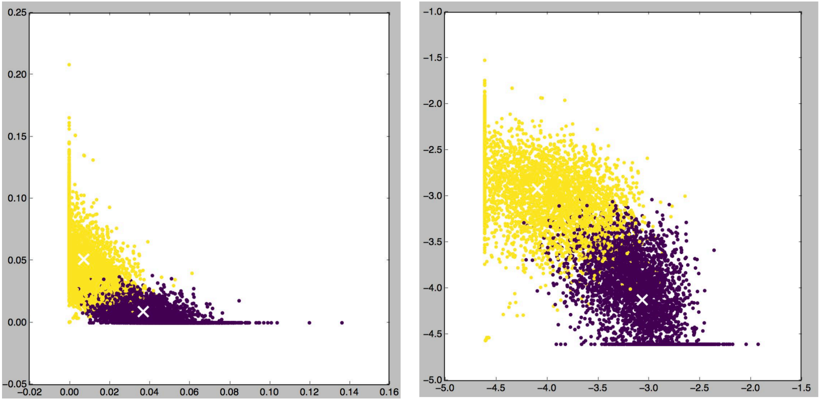


A good clustering result

Left: clustering results; Right: ground truth

Original data is high-dimensional; points are colored according to their labels

SVD is performed to project the data to 2D for visualization



The effect of logarithm transformation

Homogeneity and Completeness Score

For the purposes of the following discussion, assume a data set comprising N data points $\mathcal{U} = \{u_1, \ldots, u_N\}$, and two partitions of these:

A set of classes,

$$\mathcal{C} = \{C_1, C_2, \ldots, C_{|\mathcal{C}|}\}, \quad \bigcup_{i=1}^{|\mathcal{C}|} C_i = \mathcal{U}$$

and a set of clusters,

$$\mathcal{K} = \{K_1, K_2, \ldots, K_{|\mathcal{K}|}\}, \quad \bigcup_{i=1}^{|\mathcal{K}|} K_i = \mathcal{U}$$

Let $n_{c,k}$ be the number of data points that are members of class C_c and elements of cluster K_k:

$$n_{c,k} = |C_c \cap K_k|$$

Homogeneity and Completeness Score

Homogeneity and completeness scores are formally given by:

$$h = 1 - \frac{H(\mathcal{C} \mid \mathcal{K})}{H(\mathcal{C})}, \quad c = 1 - \frac{H(\mathcal{K} \mid \mathcal{C})}{H(\mathcal{K})}$$

where $H(\mathcal{C} \mid \mathcal{K})$ is the **conditional entropy of the classes given the cluster assignments** and is given by:

$$H(\mathcal{C} \mid \mathcal{K}) = -\sum_{k=1}^{|\mathcal{K}|} \sum_{c=1}^{|\mathcal{C}|} \frac{n_{c,k}}{N} \log \left(\frac{n_{c,k}}{|K_i|} \right)$$

and $H(\mathcal{C})$ is the **entropy of the classes** and is given by:

$$H(\mathcal{C}) = -\sum_{c=1}^{|\mathcal{C}|} \frac{|C_c|}{N} \log \left(\frac{|C_c|}{N} \right)$$

The **conditional entropy of clusters given class** $H(\mathcal{K} \mid \mathcal{C})$ and the **entropy of clusters** $H(\mathcal{K})$ are defined in a symmetric manner.

Homogeneity and Completeness Score

The **conditional entropy of clusters given class** $H(\mathcal{K} \mid \mathcal{C})$ and the **entropy of clusters** $H(\mathcal{K})$ are defined in a symmetric manner.

The **conditional entropy of clusters given class** is

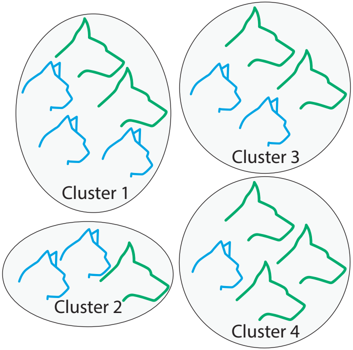
$$H(\mathcal{K} \mid \mathcal{C}) = -\sum_{c=1}^{|\mathcal{C}|} \sum_{k=1}^{|\mathcal{K}|} \frac{n_{c,k}}{N} \log\left(\frac{n_{c,k}}{|C_i|}\right)$$

and the **entropy of clusters** is:

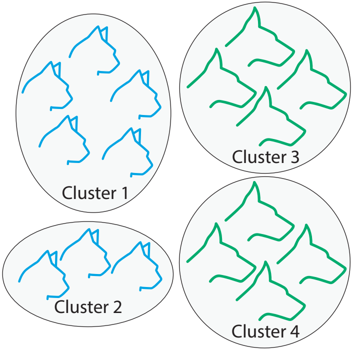
$$H(\mathcal{K}) = -\sum_{k=1}^{|\mathcal{K}|} \frac{|K_k|}{N} \log\left(\frac{|K_k|}{N}\right)$$

see also: `http://scikit-learn.org/stable/modules/clustering.html#homogeneity-completeness`

Homogeneity and Completeness Score



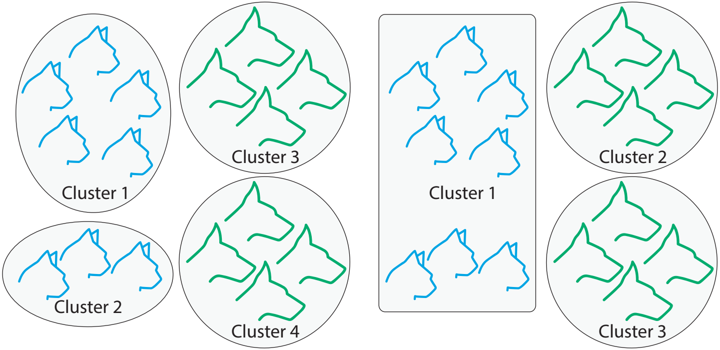
(a) Low homogeneity



(b) Perfect homogeneity

Figure 1: Homogeneity illustration

Homogeneity and Completeness Score



(a) Lower completeness (b) Higher completeness

Figure 2: Completeness illustration

V-measure

Rosenberg and Hirschberg further define **V-measure** as the harmonic mean of **homogeneity** and **completeness**:

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$