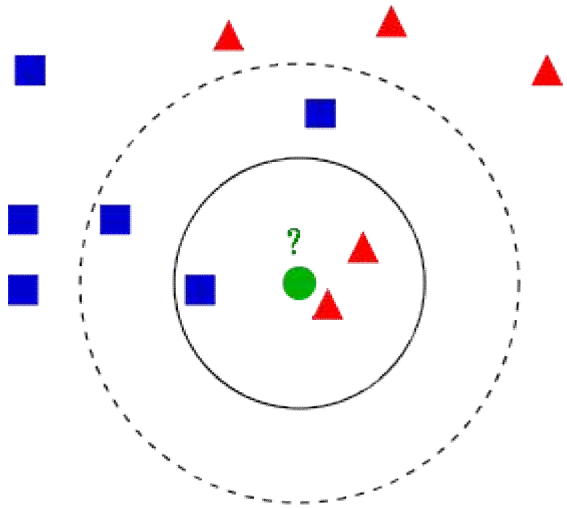
FFKNN

Feng Huang

Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

KNN introduction

In *k-NN classification*, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors.



Outline

1.Another view for KNN

2.How to search neighbors

3.Distance Metric learning For KNN

4.Extension of KNN

5. A small idea

Nonparametric and Generative model

1.KNN hasn't any assumption for data distribution.

2.How to estimate joint density distribution $p(x, \mathrm{C}_x)$ and calculate

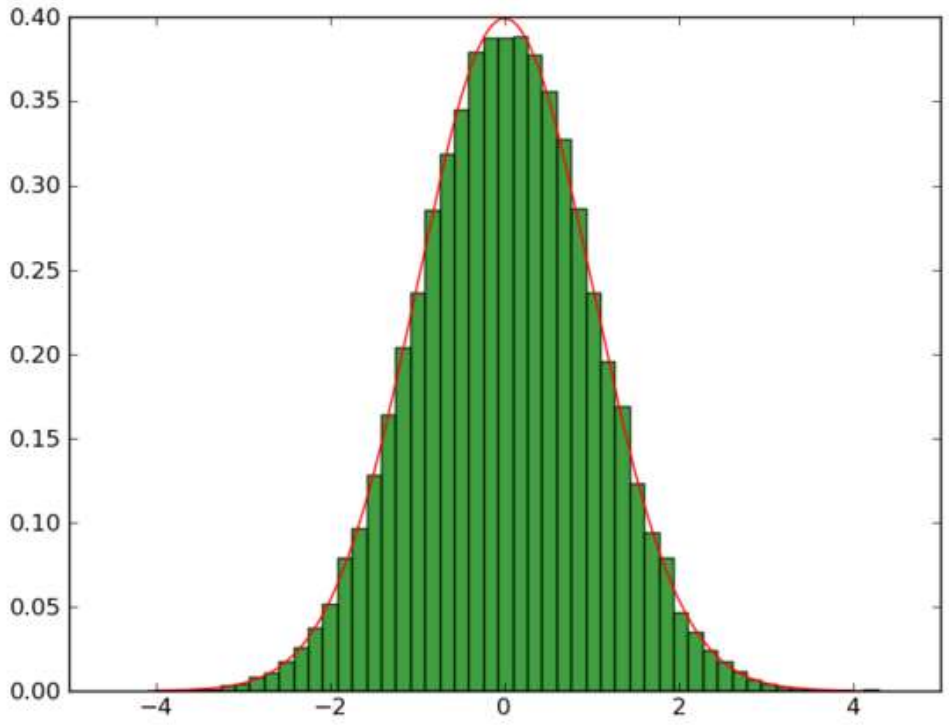
$$p(x|\mathrm{C}_x) = \frac{p(x,\mathrm{C}_x)}{p(x)}$$

Suppose this sphere has volume V and contains K_k points from

Nonparametric and Generative model

Why can we use $\frac{K_k}{NV}$ to approximate $p(x, C_x)$

We assume V is enough small:



Nonparametric and Generative model

Similarly, the unconditional density is given by

$$p(\mathbf{x}) = \frac{K}{NV}$$

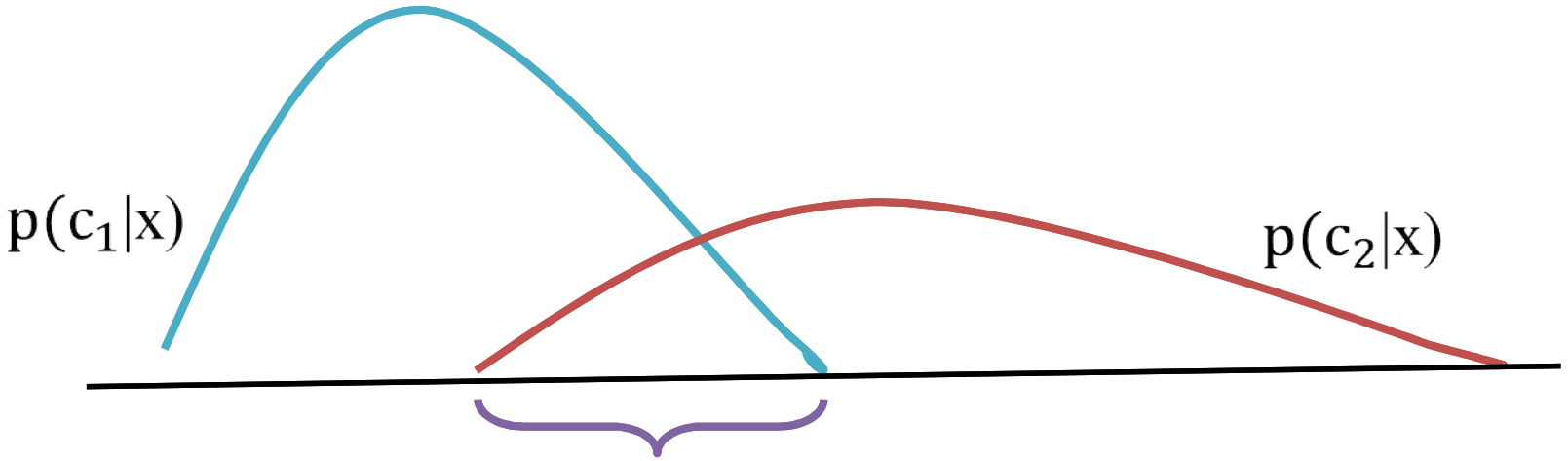
Finally, we obtain the posterior probability.

$$p(x|\mathrm{C}_x) = \frac{p(x, \mathrm{C}_x)}{p(x)} = \frac{\mathrm{K_k}}{K}$$

Error Bound of KNN

when $k = 1$, the classification error of the nearest neighbor is less than twice the Bayes probability of error.

Assume c^*=arg $max_{c \in \gamma}$P(c|x), (1 - P(c^*|x)) is the Bayes probability of error.



Error Bound of KNN

z is the nearest neighbor of x

$$P(\text{err of 1NN}) = 1 - \sum_{c\in\gamma} P(c|x)P(c|z)$$

$$\cong 1 - \sum_{c\in\gamma} P^2(c|x)$$

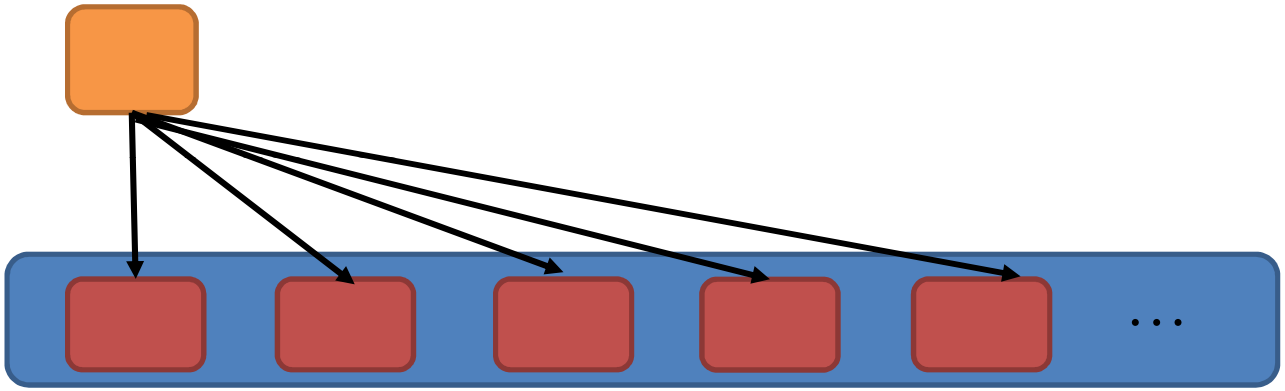
$$\leq 1 - P^2(c^*|x)$$

$$=(1+P(c^*|x))(1 - P(c^*|x))$$

$$\leq 2 \times (1 - P(c^*|x))$$

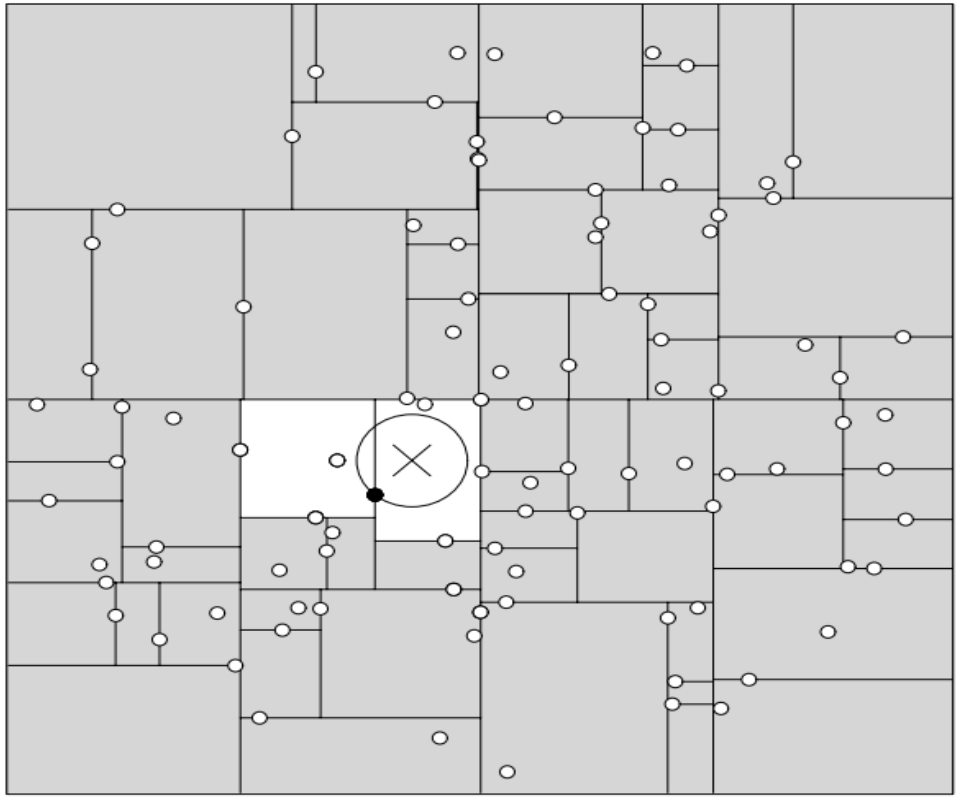
How to search neighbors

This algorithm has time complexity O(N)under the condition of Linear scanning.



KD-Tree

The basic idea is avoiding searching unnecessary feature space(data points).



Construction of KD-Tree

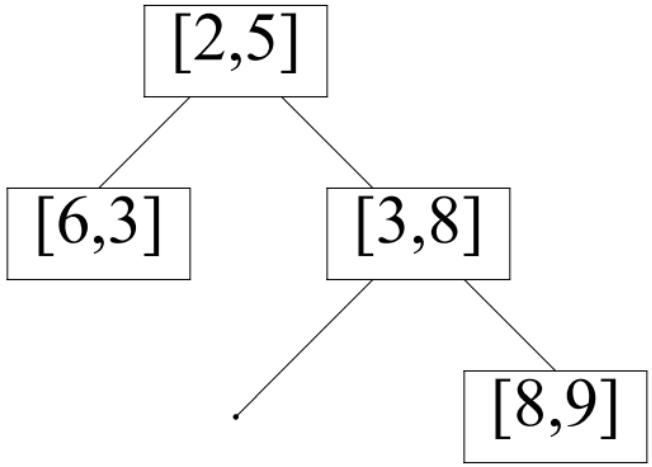
KD-Tree is a binary tree.

a) As one moves down the tree, one cycles through the axes used to select the splitting planes.

b) Points are inserted by selecting the median of the points being put into the subtree, with respect to their coordinates in the axis being used to create the splitting plane.

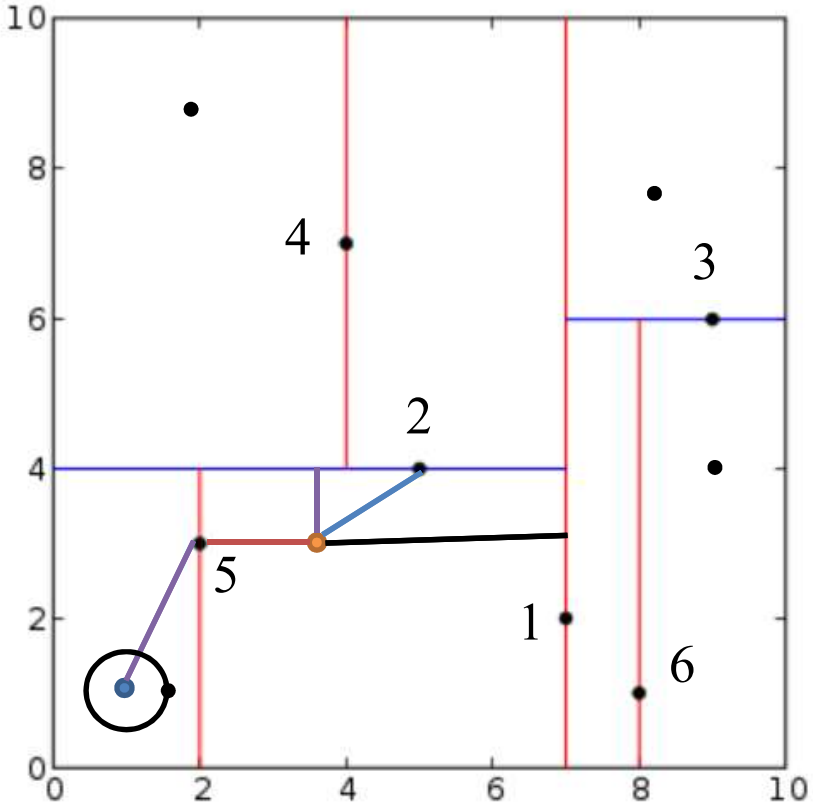
Construction of KD-Tree

A kd-tree of four elements. The splitting planes are not indicated. The [2,5] node splits along the y=5 plane and the [3,8] node splits along the x=3 plane.



How to search

For example



Distance Metric learning For KNN

The K-nearest neighbor classification performance can often be significantly improved through (supervised) metric learning.

Metric learning is the task of learning a distance function over objects. If we denote the linear transformation by a matrix A , we are effectively learning a metric $Q = A^\top A$.

$$d(x, y) = (x - y)^\top Q(x - y) = (Ax - Ay)^\top (Ax - Ay)$$

Neighbourhood Components Analysis

The target of NCA is optimizing leave-one-out (LOO) performance on the training data. The actual leave-one-out classification error of KNN is quite a discontinuous function.

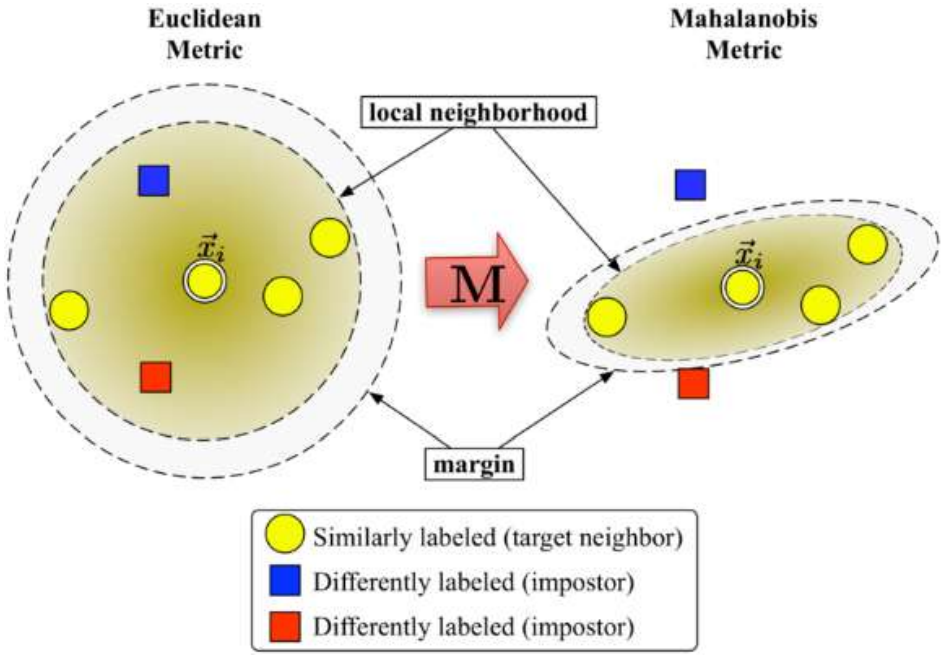
$$p_{ij} = \frac{\exp(-\|Ax_i - Ax_j\|^2)}{\sum_{k \neq i} \exp(-\|Ax_i - Ax_k\|^2)} \quad , \quad p_{ii} = 0$$

denote the set of points in the same class as i by $C_i = \{j | c_i = c_j\}$.

$$p_i = \sum_{j \in C_i} p_{ij}$$

Large Margin Nearest Neighbor

The metric is optimized with the goal that k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin.



Large Margin Nearest Neighbor

Object function is

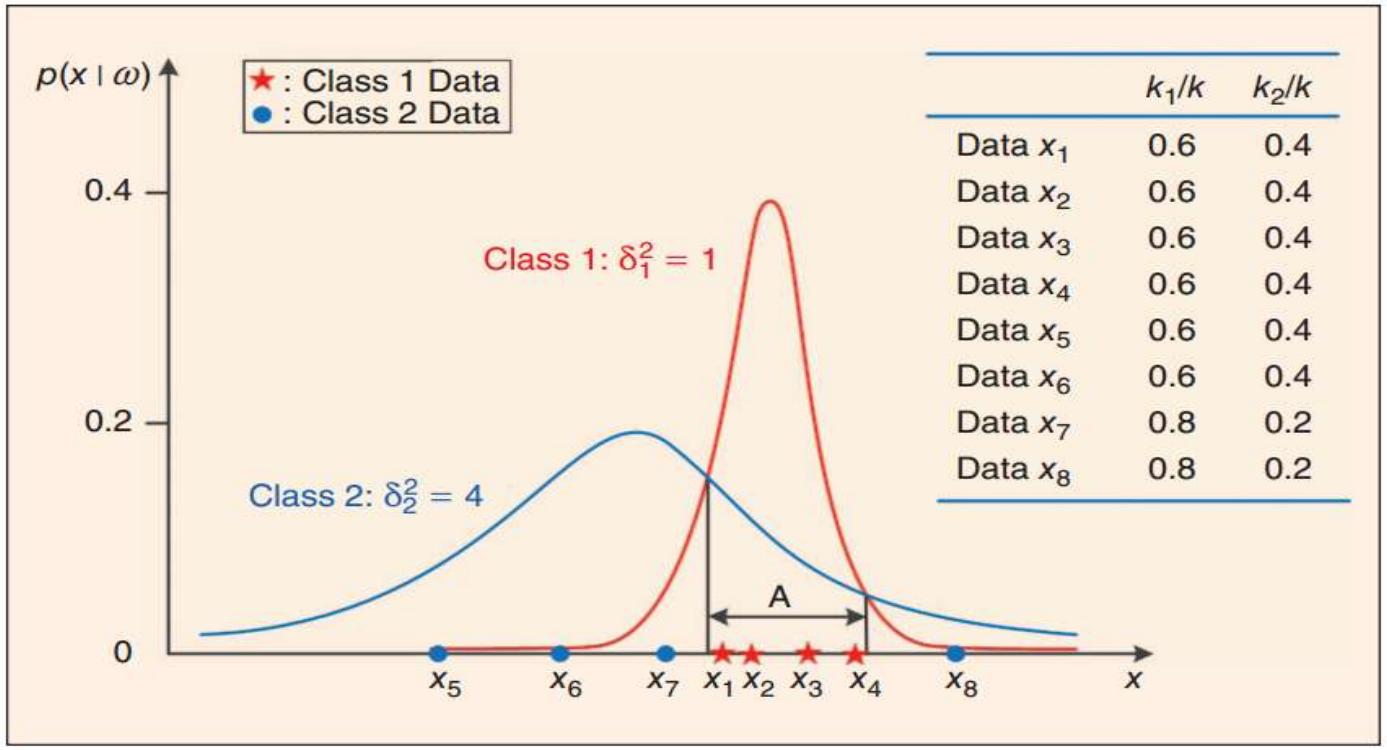
$$\varepsilon(\mathbf{L}) = \sum_{ij} \eta_{ij} \|\mathbf{L}(\vec{x}_i - \vec{x}_j)\|^2 + c \sum_{ijl} \eta_{ij}(1 - y_{il}) \left[1 + \|\mathbf{L}(\vec{x}_i - \vec{x}_j)\|^2 - \|\mathbf{L}(\vec{x}_i - \vec{x}_l)\|^2\right]_+$$

The first term penalizes large distances between each input and its target neighbors, while the second term penalizes small distances between each input and all other inputs that do not share the same label.

In my opinion, the objection function just likes optimizing many local SVM.

Extension of KNN: ENN

It considers not only who are the nearest neighbors of the test sample, but also who consider the test sample as their nearest neighbors.



	k_1/k	k_2/k
Data x_1	0.6	0.4
Data x_2	0.6	0.4
Data x_3	0.6	0.4
Data x_4	0.6	0.4
Data x_5	0.6	0.4
Data x_6	0.6	0.4
Data x_7	0.8	0.2
Data x_8	0.8	0.2

ENN

Define the generalized class-wise statistic T_i for class i as the following:

$$T_i = \frac{1}{n_i k} \sum_{\mathbf{x} \in S_i} \sum_{r=1}^{k} I_r(\mathrm{x}, S = S_1 \cup S_2)$$

$$i = 1, 2$$

where

$$I_r(\mathrm{x}, S) = \begin{cases} 1, & \text{if } \mathrm{x} \in S_i \text{ and } \mathrm{NN}_r(\mathrm{x}, S) \in S_i \\ 0, & \text{otherwise} \end{cases}$$

ENN

Given an unknown sample Z to be classified, we iteratively assign it to class 1 and class 2, respectively, to obtain two new generalized class-wise statistics.

The ENN classifier predicts its class membership according to the following target function:

$$f_{\mathrm{ENN}} = \arg\max_{j \in 1,2} \sum_{i=1}^{2} T_i^j$$

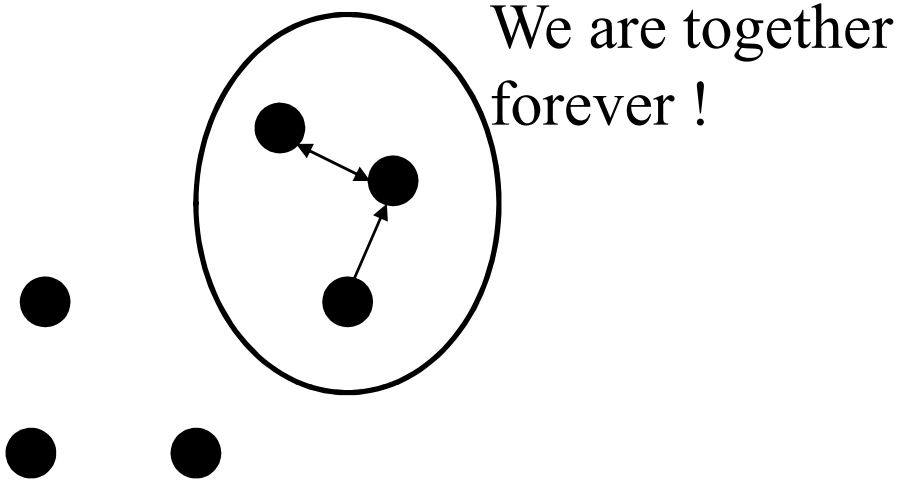
K-Nearest-Neighbor Consistency

Incorporating Local Information into Global Optimization

> **Cluster k-Nearest-Neighbor Consistency**: For any data object in a cluster, its k-nearest neighbors should also be in the same cluster.

The paper presents an algorithm to enforce 100% kNN consistency of the clustering results from a standard clustering algorithm such as the K-means algorithm.

Closed neighbor set



K-Nearest-Neighbor Consistency

Initialize cluster centers $(c_1, \cdots, c_K)$

Iterate (1) and (2) until converge:

1. Assign cluster membership. Assign one closed-neighbor set S at a time. Assign all objects of the closed-neighbor set S to the closest cluster C_p , where the closeness is defined in average sense:

$$p = \arg\min_k \sum_{i \in S} (\mathbf{x}_i - \mathbf{c}_k)^2$$

2. Update centers: $c_k = \sum_{i \in C_k} x_i / n_k$

My Idea

迭代进行一下步骤，直到max次，或正确率大于某一阈值。

1.根据上一轮分类器的错误情况，赋予数据权重，进行有放回的抽样。

2.采用同步聚类对采样数据进行聚类（怎么修改？）。

3.聚类得到的数据点用于本轮KNN分类器的构建。

最后这n轮迭代的KNN分类器，其实就是每轮最后的聚类中心点应该怎么组合的问题？

Take home message

1. Learn KNN from the view of density estimation.
2. Split feature space to search nearest neighbor fast.
3. Metric learning can find suitable distance measure.
4. Locality of KNN

Thanks