K-Nearest Neighbors（K最邻近）

- 一般用于分类问题

案例：

已知数据 $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$

对应的标签为 $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$，$y_i \in \{1, 2, 3, \ldots, C\}$ 其中 $i=1, \ldots, n$.

判断输入数据 x_0 的类别，其中 $x_0 = (x_{01} \ x_{02} \ \cdots \ x_{0d})$

KNN 算法：

1. 计算 x_0 与 X 中每个向量的距离（欧氏距离 $d(u, v) = \|u - v\|_2 = \sqrt{\sum_{i=1}^{d} (u_i - v_i)^2}$）

$$dist = [\|x_0 - x_1\|_2, \|x_0 - x_2\|_2, \ldots, \|x_0 - x_n\|_2]$$

2. 从 dist 中取前 K 小的元素，说明 x_0 与这 K 个元素对应的向量距离最近

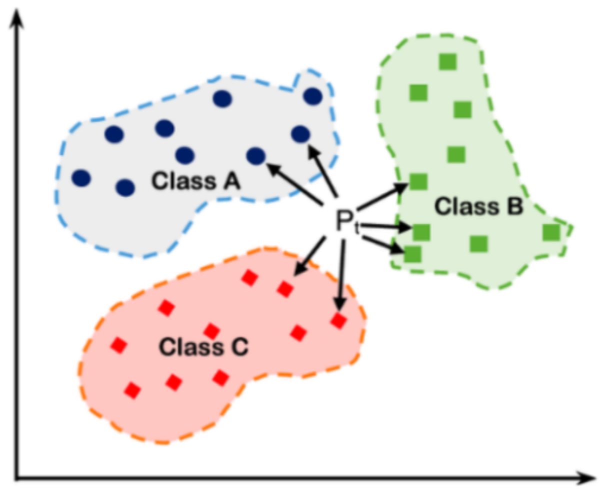
K个最近向量的标签为：$c_1, c_2, \ldots, c_K \in \{1, 2, \ldots, C\}$

这K个标签进行投票，得票多的标签即为 x_0 的标签。

如果有平票可以随机选择一个.

可视化解释

K Nearest Neighbors



如何同时计算多个数据类别？

计算 $t_1, t_2, \ldots, t_T$ 和 X 的距离，其中 $t_i=(t_{i1}, t_{i2}, \ldots, t_{id})$ $1\le i\le T$

同上可知 $dist(t_1, X) = [\|t_1-x_1\|_2, \|t_1-x_2\|_2, \ldots, \|t_1-x_n\|_2]$

$$dist(t_2, X) = [\|t_2-x_1\|_2, \|t_2-x_2\|_2, \ldots, \|t_2-x_n\|_2]$$

$$\vdots$$

$$dist(t_T, X) = [\|t_T-x_1\|_2, \|t_T-x_2\|_2, \ldots, \|t_T-x_n\|_2]$$

$$dist = \begin{bmatrix} dist(t_1, X) \\ dist(t_2, X) \\ \vdots \\ dist(t_T, X) \end{bmatrix} = \begin{pmatrix} \|t_1-x_1\|_2, & \|t_1-x_2\|_2, & \ldots, & \|t_1-x_n\|_2 \\ \|t_2-x_1\|_2, & \|t_2-x_2\|_2, & \ldots, & \|t_2-x_n\|_2 \\ \vdots & \vdots & & \vdots \\ \|t_T-x_1\|_2, & \|t_T-x_2\|_2, & \ldots, & \|t_T-x_n\|_2 \end{pmatrix}$$

记 $dist^2 = \begin{pmatrix} \|t_1 - x_1\|_2^2, & \|t_1 - x_2\|_2^2, & \dots, & \|t_1 - x_n\|_2^2 \\ \|t_2 - x_1\|_2^2, & \|t_2 - x_2\|_2^2, & \dots, & \|t_2 - x_n\|_2^2 \\ \vdots & \vdots & & \vdots \\ \|t_T - x_1\|_2^2, & \|t_T - x_2\|_2^2, & \dots, & \|t_T - x_n\|_2^2 \end{pmatrix}$

$$= \begin{pmatrix} \|t_1\|_2^2 + \|x_1\|_2^2 - 2t_1 \cdot x_1^T, & \|t_1\|_2^2 + \|x_2\|_2^2 - 2t_1 \cdot x_2^T, & \dots, & \|t_1\|_2^2 + \|x_n\|_2^2 - 2t_1 x_n^T \\ \|t_2\|_2^2 + \|x_1\|_2^2 - 2t_2 \cdot x_1^T, & \|t\|_2^2 + \|x_2\|_2^2 - 2t_2 \cdot x_2^T, & \dots, & \|t_2\|_2^2 + \|x_n\|_2^2 - 2t_2 x_n^T \\ \vdots & \vdots & & \vdots \\ \|t_T\|_2^2 + \|x_1\|_2^2 - 2t_T x_1^T, & \|t_T\|_2^2 + \|x_2\|_2^2 - 2t_T \cdot x_2^T, & \dots, & \|t_T\|_2^2 + \|x_n\|_2^2 - 2t_T x_n^T \end{pmatrix}$$

$$= \begin{pmatrix} \|t_1\|_2^2 & \|t_1\|_2^2 & \dots & \|t_1\|_2^2 \\ \|t_2\|_2^2 & \|t_2\|_2^2 & \dots & \|t_2\|_2^2 \\ \vdots & \vdots & & \vdots \\ \|t_T\|_2^2 & \|t_T\|_2^2 & \dots & \|t_T\|_2^2 \end{pmatrix}_{T\times n} + \begin{pmatrix} \|x_1\|_2^2 & \|x_2\|_2^2 & \dots & \|x_n\|_2^2 \\ \|x_1\|_2^2 & \|x_2\|_2^2 & \dots & \|x_n\|_2^2 \\ \vdots & \vdots & & \vdots \\ \|x_1\|_2^2 & \|x_2\|_2^2 & \dots & \|x_n\|_2^2 \end{pmatrix}_{T\times n} - \begin{pmatrix} 2t_1 \cdot x_1^T & 2t_1 \cdot x_2^T & \dots & 2t_1 x_n^T \\ 2t_2 \cdot x_1^T & 2t_2 \cdot x_2^T & \dots & 2t_2 x_n^T \\ \vdots & \vdots & & \vdots \\ 2t_T \cdot x_1^T & 2t_T x_2^T & \dots & 2t_T x_n^T \end{pmatrix}$$

(broadcast)
广播机制

$$= \begin{pmatrix} \|t_1\|_2^2 \\ \|t_2\|_2^2 \\ \vdots \\ \|t_T\|_2^2 \end{pmatrix}_{T\times 1} + \begin{pmatrix} \|x_1\|_2^2 & \|x_2\|_2^2 & \dots & \|x_n\|_2^2 \end{pmatrix}_{1\times n} - 2 \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_T \end{pmatrix}_{T\times d} \begin{pmatrix} x_1^T & x_2^T & \dots & x_n^T \end{pmatrix}_{d\times n}$$