

MLPR第一讲学习笔记

by zijeff

模式（样本）的表示方法：

- **向量表示**：假设一个样本有 n 个变量或者特征，我们很自然会用一个 n 维向量来存储这个样本的变量或者特征。

$$\mathbf{x} = (x_1, x_2, x_3, \dots, x_n)^T$$

- **矩阵表示**：那如果有多个样本，比如 m 个样本，也就是有 m 个 n 维向量。显而易见的，我们可以用一个 $m \times n$ 的**矩阵**来表示。

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{pmatrix}$$

- **坐标表示**：我们高中都学过，任何一个向量都可以在一个线性空间里来表示。比如说一维向量可以表示数轴上的点，二维向量可以表示平面上的点，三维向量可以表示空间上的点.....更高维的向量对我们来说就不是很直观了，这也就是**样本的坐标表示**。
- **基元（链码）表示**：定义方向和基元线段长度，在**句法模式识别**中会用到。

模式类的紧致类：

- **紧致集**：同一类模式类样本的分布比较集中，没有或临界样本很少，这样的模式类称**紧致集(compact set)**。针对这个概念，我们可以运用矩阵来举个例子，方便直观地理解。我们不妨设有 3×3 这样的**一个方阵**，里面排列的事物**只有0和1这两种属性**。例如下面表格中给出的三种排列，前两个矩阵的排列一眼可以认为**紧致集**，具有不同特征的样本较为**集中**分布。而最后一个矩阵中具有两种属性的样本交替出现，不满足紧致集。当满足紧致集的时候，才能很好的分类；如果不满足紧致集，就要采取变换的方法，以满足紧致集。

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

相似与分类：

- **样本之间相似度量要满足以下需求**：应为非负值；样本本身相似性度量应最大；度量应满足对称性；在满足紧致性的条件下，相似性应该是点间距离的单调函数。
- **常用各种距离表示样本相似性**：不妨设有两个具有 n 个特征的样本，将其记为 \mathbf{x}_i 和 \mathbf{x}_j ，向量表示如下：

$$\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i2}, \dots, x_{in})^T$$

$$\mathbf{x}_j = (x_{j1}, x_{j2}, x_{j2}, \dots, x_{jn})^T$$

1. 绝对值距离:

$$d(i, j) = \sum_{k=1}^n |x_{ik} - x_{jk}|$$

2. 欧几里得距离:

$$d(i, j) = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}$$

3. 闵可夫斯基距离:

$$d(i, j) = \left(\sum_{k=1}^n |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}}$$

显然地, 当 $p = 1$ 时, 上式变为绝对值距离; $p = 2$ 时, 上式变为欧几里得距离。

4. 切比雪夫距离:

$$d(i, j) = \max_k |x_{ik} - x_{jk}|$$

5. 马氏距离:

$$d(i, j) = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)}$$

其中, S 为协方差矩阵。

6. 夹角余弦:

$$\cos \theta = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} = \frac{\sum_{k=1}^n x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^n x_{ik}^2} \sqrt{\sum_{k=1}^n x_{jk}^2}}$$

很显然, 夹角余弦就是我们在高中数学中所接触到的算两个向量的夹角。从几何平面上直观想象一下, 两个向量夹角越小, 它们就靠得越近, 可以认为这两个向量的相似程度很高。

7. 相关系数:

$$\rho = \frac{\sum_{k=1}^n (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^n (x_{jk} - \bar{x}_j)^2}}$$

这个公式也是我们在高中数学中计算线性回归时会使用的公式之一, 用来反映相关性。

特征的生成：

- 低层特征：
 1. 无序尺度：有明确的数量和数值。
 2. 有序尺度：有先后，好坏这种可以比较的次序关系。
 3. 名义尺度：无数量，次序关系。
- 中层特征：经过计算和变换后得到的特征。
- 高层特征：在中层特征的基础上有目的的经过运算形成。