# Part III-B: Medicine AI

Lecture by None

Note by THF

2024 年 10 月 23 日

# 目录



## Learn 4

10.20

## 0.1 数据预处理

### 0.1.1 标准化

**Notation.** 变量离差标准化：标准化后所有变量范围都在 [0,1] 内

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}.$$

**Example.** 一组变量如下：

$$X = (1.5, 1.7, 2.2, 1.2, 1.6, 1.4, 1.1)\,.$$

易得 $x_{\min}=1.1, x_{\max}=2.2$

$$\begin{aligned} y_i &= \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \\ &= \frac{x_i - 1.1}{2.2 - 1.1} \\ &= \frac{x_i - 1.1}{1.1} \\ &= \frac{x_i}{1.1} - 1 \end{aligned}$$

.

得 $Y = (0.364, 0.545, 1, 0.091, 0.455, 0.273, 0)$

**Notation.** *Z-score*（变量标准差）标准化

经过标准化后平均值为 0，标准差为 1

$$z_i = \frac{x_i - \bar{x}}{s} \quad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

可以看出 $s$ 为原数据的标准差，$z_i$ 值其实等同于标准正态分布中的 $u$ 值：

$$u = \frac{x - \mu}{\sigma} \quad y = \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-u^2}/2.$$

### 0.1.2 插补缺失值

**Notation.** 均值插补

1. 数值性变量：采用平均值插补
2. 离散型：采用众数插补

**Notation.** 同类均值插补：使用层次聚类方法归类缺失值的样本，用该类别的特征均值插补

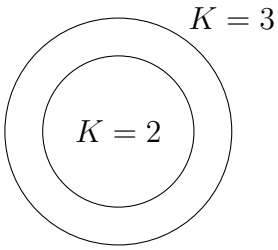
**Notation.** KNN(*K-nearest neighbor*) 缺失值插补：找到与含缺失值样本相似的 $K$ 个样本，使用这 $K$ 个样本在该缺失变量上的均值填充

# *K-nearest neighbor*

## 基本思路

找到与新输入的待预测样本最临近的 $K$ 个样本，判断这 $K$ 个样本中绝大多数的所属类别作为分类结果输出

条件：已经具有较大的样本量




**Notation.** KNN 算法的基本要素：距离度量、$K$ 值、分类决策规则

## 距离度量

**Notation.** KNN 算法能够分类：特征空间内的样本点之间的距离能够反映样本特征的相似程度

设有两个样本点 $\boldsymbol{x}_i, \boldsymbol{x}_j$ ，以 $n$ 维向量空间作为特征空间，将这两个点表示为：

$$\boldsymbol{x}_i, \boldsymbol{x}_j \in \boldsymbol{X}.$$

$$\boldsymbol{x}_i = \left(x_i^1, x_i^2, \ldots, x_i^n\right)^T.$$

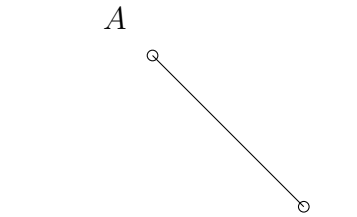
$$\boldsymbol{x}_j = \left(x_j^1 . x_j^2, \ldots, x_j^n\right)^T.$$

特征点之间的距离定义为：

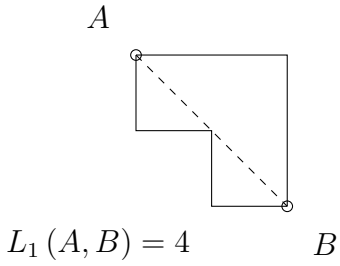
$$L_p\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\sum_{l=1}^{n} \left|x_i^l - x_j^l\right|^p\right)^{\frac{1}{p}}.$$

**Example.** 代入 $p=2$ ，易得 $L_2\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$ 为平面上两点间的距离公式，该距离又称为欧氏距离：

$$L_2\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \sqrt{\left(x_{i_1} - x_{j_1}\right)^2 + \left(x_{i_2} - x_{j_2}\right)^2}.$$




代入 $p=1$： $L_1(x_i,x_j)$ 称为曼哈顿距离：




## $K$ 值的选择

使用交叉验证方法确定最合适的 $K$ 值

## Learn 5

10.23

## 0.2 模型评估和性能度量

**Notation.** 留出法 (*hold-out*)：

将原始数据集 $D$ 分为两个互斥的子集 $S,T$，$S$ 作为训练数据集，$T$ 作为测试数据集：$D=S\cup T, S\cap T=\varnothing$

在划分任务时要尽量保证 $S$ 和 $T$ 中的样本类别比例相似

**Example.**

$$D(a,b)\to S(\lambda a,\lambda b)\cup T((1-\lambda)a,(1-\lambda)b).$$

该过程称为分层采样法，其中 $\lambda\in[\frac{2}{3},\frac{4}{5}]$

使用 $S$ 训练模型，$T$ 进行模型测试，多次随机划分 $a,b$ 在 $S$ 和 $T$ 内的内容，多次实验取测试结果平均值

**Notation.** 交叉验证法/$k$ 折交叉验证 (*cross validation*/*k-fold cross validation*):

$$D = D_1 \cup D_2 \cup \ldots \cup D_k \text{ 且} D_I \cap D_j = \varnothing \, (i \neq j) \, .$$

此处 $\forall D_i$ 由 $D$ 分层采样得到

每次实验使用 $k-1$ 个子集的并集训练，剩下的一个子集作为测试集:

$$S = \sum_{i=1}^{m-1} D_i + \sum_{i=m+1}^{k} D_i \quad T = D_m.$$

取不同的 $m$ 值共可以得到 $k$ 组 “训练集-测试集”，得到 $k$ 个结果，取 $k$ 个结果的平均值

**Example.** 5 折交叉验证的数据划分:

| | | | | |
|---|---|---|---|---|
| $D_1$ | $D_2$ | $D_3$ | $D_4$ | $\boldsymbol{D}_5$ |
| $D_1$ | $D_2$ | $D_3$ | $\boldsymbol{D}_4$ | $D_5$ |
| $D_1$ | $D_2$ | $\boldsymbol{D}_3$ | $D_4$ | $D_5$ |
| $D_1$ | $\boldsymbol{D}_2$ | $D_3$ | $D_4$ | $D_5$ |
| $\boldsymbol{D}_1$ | $D_2$ | $D_3$ | $D_4$ | $D_5$ |

$$\Rightarrow \begin{cases} Res_1 \\ Res_2 \\ Res_3 \\ Res_4 \\ Res_5 \end{cases} \xrightarrow{\text{Avg}\cdot} \text{Result}$$

**Notation.** 若样本量 $m$ 等于子集数 $k$ ，交叉验证法等同于留一法 (*leave one out*, LOO)

留一法的优点：训练结果更准确

缺点：样本量太大的时候消耗过多资源

## 0.3 模型性能度量

**Notation.** 错误率:

$$E = \frac{1}{m} N \left( f \left( x_i \right) \neq y_i \right) .$$

准确率:

$$\text{Acc} = \frac{1}{m} N \left( f \left( x_i \right) = y_i \right) .$$

$m$ 为样本总数，$N \left( f \left( x_i \right) = y \right)$ 表示符合特征 $f : x \to y$ 的样本数量