INT104W2_数据预处理

> *"数据质量有时往往比算法更加重要"*

数据集

数据类型：结构化（如表格），非结构化（如自然语言文本，网页）

几种常见数据格式：

- CSV(Comma Separated Values)

  Null,13,6,8

- TSV(Tab Separated Values)

  Name<Tab>Age<Tab>Adress

- XML (Extensible Markup Language)

```
<\?xml version\= ……
```

- JSON (JavaScript Object Notation)

```
{
"name": "Jack",
"gender": "male"
}
```

几种Python图形/数据可视化库：Matplotlib，Seaborn，**Pandas**.plot

Histograms：直方图

Scatter Plots：散点图

Heat Maps：热力图

数据预处理（Data Pre-processing）：

Data Cleaning数据清洗

- 解决丢失的数据（missing data）：可以去除整行数据；去除整列数据（如果大部分缺失）；使用平均值，中位数，0等填入缺失数据

- 平滑噪声数据（没有办法完全去除或平滑数据中的噪声）：可以删去，也可以尝试将其恢复为正确值
- 噪声：错误的信息，可以是不符合实际意义的（负数），可以是超出正常值过多的（需要证实）

总的来说，结果清洗，数据的**行数和列数都会减少**

Data Integration数据集成

- Combine合并：将多个来源的数据合并到一个文件（位置）
- Resolve conflicts解决冲突：如转化英制公制单位
- Remove redundant删除冗余：同一类别的数据在不同数据来源中可能有不同名字，需将其合并（相关性分析）

Data Transformation数据转换

使数据保持一致性和可读性

几种编码方式

- Ordinal encoder有序编码：[1][2][3][4]
- One-hot encoder独热编码:[1,0,0,0][0,1,0,0]

> 机器学习—特征工程—OneHotEncoder独热编码
>
> https://blog.csdn.net/weixin_45252110/article/details/98749238

几种标准化/归一化/中心化方法与特性：

- Min–max normalization：

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Z-score** normalization标准分数：使所有值的平均值为0，标准差为1（即将原坐标原点变换为数据集重心，实现中心化）

$$X_{scaled} = \frac{X - X_{mean}}{X_{std}}$$

- Normalization by decimal scaling：用十进制归一化

$$X_{scaled} = \frac{X}{10^j}$$

【机器学习】数据归一化全方法总结：Max-Min归一化、Z-score归一化、数据类型归一化、标准差归一化等

https://blog.csdn.net/Next_SummerAgain/article/details/127321209

Data Reduction数据缩减

在这个过程中，能获得相似结果的数据集将被简化表示（减少行/列）

Feature Selection特征选择

Filter methods过滤式：

数据将被按照他们与目标的关系选择并排序，**先对数据集进行特征选择，然后再训练模型，特征选择过程与后续模型训练无关**，**Filter方法常用的特征子集评价标准包括：相关系数**、互信息、信息增益等，更多方法参见 mlr 包支持的所有 Filter 方法

Wrapper methods包裹式：

包裹式特征选择直接把最终将要使用的模型的性能作为特征子集的评价标准，也就是说，包裹式特征选择的目的就是为给定的模型选择最有利于其性能的特征子集，LVM（Las Vegas Wrapper）是一个典型的包裹式特征选择方法

Embedded methods嵌入式：

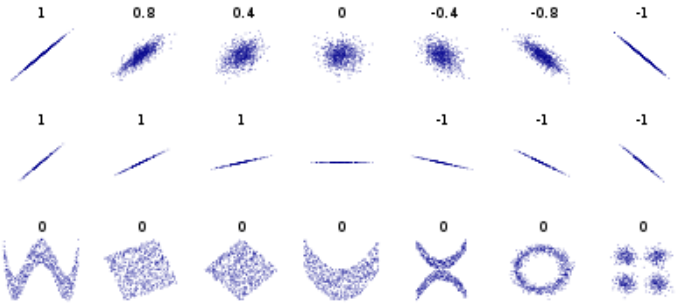
在前两种特征选择方法中，特征选择过程和模型训练过程是有明显分别的两个过程，嵌入式特征选择是将特征选择过程与学习器训练过程融为一体，两者在同一个优化过程中完成，即在学习器训练过程中自动地进行了特征选择。如 LASSO 和 Ridge Regression

Looking for Correlations寻找相关性

相关性是一种统计分析，用于衡量和描述两个变量之间关系的强度和方向。

皮尔逊相关系数Pearson Correlation：衡量线性相关性

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$



注：相关信息见W3笔记

Feature Extraction特征提取

从现有特征中提取新特征，从数据集中识别和选择最相关和最有用的特征，将它们**转换为低维空间**，同时**保留最重要的信息。**

常见方法：

PCA(Principal Component Analysis)主成分分析

CNN(Convolutional Neural Networks)卷积神经网络

请自行拓展