数据预处理

该课程主要为大家讲授如下的内容：

- 数据预处理（缺失值插补、去重、特征选择、标准化、标签转化）

1. 数据预处理
    1. 缺失值插补
    单变量缺失值插补方法SimpleImputer
    初始化代码样式为：
    `imp_mean = SimpleImputer(missing_values=np.nan, strategy='mean', copy=True)`
    该方法中需要注意的参数有：
    missing_values：缺失值的占位符，即该标志出现即表示此处缺失。
    Strategy：缺失值的插补策略{“mean”, “median, “most_frequent”, “constant”}
    分别表示：均值、中位数、众数、常量。
    选择“constant”时，需要附加参数fill_value，来填入参数。
    Copy：为true是表示创建副本返回，为false时表示在原数据上进行插补。 该方法是一个类，实际使用中还需要通过相应的接口（API）来使用：
    fit(X)：对输入数据进行运算，根据插补策略确定每个特征分量的插补值。
    transform(X)：对输入数据进行插补操作，将通过fit方法计算出的插补值插补到缺失值上
    fit_transform(X)：对数据进行运算，并将缺失值插补
    2. 去重
    数据去重，这个一般是表格数据才会用到，可以通过pandas的去重方法duplicated来完成。
    代码样式为：
    `data.drop_duplicates(subset=['A','B'],keep='first',inplace=True)`
    该方法中需要注意的参数有：

Subset：列名，以该列数据为准去重，默认为none
Keep：保留策略，默认为false
first：保留第一次出现的重复行
last：保留最后一次出现的重复行
False：删除所有重复行.
Inplace：操作策略，默认为false
true：在原数据上进行操作
false：删除重复行后返回一个副本

3. 特征选择
方差过滤：设定方差阈值，方差低于此值的特征将被删除。
代码样式为：
`sklearn.feature_selection.VarianceThreshold(threshold=0.0)`
该方法中需要注意的参数有：
threshold：方差阈值，方差低于此值的特征将被删除

4. 特征提取
分词统计，在类CountVectorizer中，实现了词语切分和词语出现次数统计。
代码样式为：
`CountVectorizer(input='content', encoding='utf-8‘, lowercase=True)`
该方法中需要注意的参数有：
input：输入的内容形式，默认为content
filename：原始文件名
file：通过read函数读入的数据
content：字符串或字节类型
Encodeing：数据的编码方式，默认为utf-8
Lowercase：将所有字符转为小写后再进行操作 数据标准化：
标准化：`preprocessing.StandardScaler(X)`
最大最小化：`preprocessing.MinMaxScaler(feature_range=(0, 1),X)`
eature_range：所需转换的数据范围
标签转换，用于变换监督学习的目标.
标签二值化：在机器学习处理过程中，为了方便算法的实现，经常需要把字符串的标签数据转化成整数索引。

`sklearn.preprocessing.LabelBinarizer()`

标签转换：根据标签位置就可以把文字特征转换成数字特征。

`sklearn.preprocessing.LabelEncoder()`