一、数据清洗

    1.1 预处理

    1）数据处理工具的选择： 一般选用数据库处理和Python等工具处理。

    2）查看数据的元数据以及数据特征

    1.2 缺省值处理（删除、补全）

     缺省值处理的一般步骤：确定缺省值范围、去除不需要的字段、填充缺省值内容（经验值、均值、中位数、众数、推测值）、重新获取数据。

    1.3 格式、内容错误的数据（修改、删除）

    1.4 逻辑错误的数据（修改、删除）

    1.5 不需要的数据（删除）

    1.6 关联性验证

二、文本数据处理

    1、词袋法、词集法

    1）词袋法。词袋法计算文档中每个单词出现的次数（词频）。

    2）词集法。词集法是判断单词是否出现在文档中，出现就为1，否则为0。它是考虑到文档次数相差太大的时候，词袋法不适用的情况。

    2、TF-IDF（词频-逆文档频率）

    1）词频-逆文档频率。

    TF-IDF认为：单词的重要性随着它在文本中出现的次数成正比增加，也就是单词的出现次数越多，该单词对于文本的重要性就越高。同时单词的重要性会随着在语料库中出现的频率成反比下降，也就是单词在语料库中出现的频率越高，表示该单词与常见，也就是该单词对于文本的重要性越低。

    2）计算公式

    TF-IDF=TF\*IDF

    其中，TF=文档中词出现的次数/文档中词的总个数；IDF=log（总的文档个数/出现该词的文档数量+1），这里的-1操作主要是防止分母为0（也就是说出现该词的文档数量为0）

    3）WordToVictor（直接转化为词向量）

三、数值型数据处理（sklearn.Preprocessing）

    3.1 缺省值填充

      Imputer(missing\_values='NaN',strategy='mean',axis=0),其中axis=0表示按照列进行填充

      对于缺省的数据，在处理之前一定需要进行预处理操作，一般采用中位数、均值或者众数来进行填充，在scikit中主要通过Imputer类来实现对缺省值的填充。

    3.2 哑编码(OneHotEncoder)--one of K

       哑编码：对于定性的数据(也就是分类的数据)，可以采用N位的状态寄存器来对N个状态进行编码，每个状态都有一个独立的寄存器位，并且在任意状态下只有一位有效。

       哑编码是一种常用的将特征数字化的方式。比如有一个特征属性:['male','female']，那么male使用向量[1,0]表示，female使用[0,1]表示。

    3.3 二值化(Binarizer(threshold=0.0))

    对于给定的数据，通过给定阈值，将所有大于阈值的数据变为1，所有小于阈值的数据变为0。

    3.4 标准化

    StandardScale(copy=True, with\_mean=True, with\_std=True)

    基于特征属性的数据(也就是特征矩阵的列)，获取均值和方差，然后将特征值转换至服从标准正态分布的值。

    其计算公式为：,其中表示标准化后的值，x表示原始数据，表示特征矩阵的列的均值，表示特征矩阵的列的均值。

    3.5 区间缩放法(等比例缩放）

    MinMaxScaler(feature-range=(0,1),copy=True)，其中feature-range为要缩放的数据的范围。

    按照数据的方差特性将数据缩放到给定区间上。

    3.6 归一化（行处理）

    Normalizer（norm=‘l2’，copy=True）

    将矩阵的行转化为单位向量（行值相加为1）。

    3.7 多项式扩展

    PolynomialFeatures(degree=2, interaction\_only=False, include\_bias=True)

    所谓多项式扩展，就是按照特定的多项式规则将数据集的每一个数据的特征进行相应阶数的扩展。

四、特征选择

    4.1 特征选择的两个考虑

    1）特征是否发散（越发散越好，越发散越能够达到模型训练的目的）

    2）特征与目标的相关性（相关性越高越好）

    4.2 特征选择的方法

    1）过滤法（Filter）

    过滤法，按照发散性或者相关性对各个特征进行评分，设定阈值或者待选择阈值的个数，从而选择特征；常用方法包括方差选择法、相关系数法、卡方检验、互信息法等。

    2）包装法（Wrapper）

    包装法，根据目标函数（通常是预测效果评分），每次选择若干特征或者排除若干特征；常用方法主要是递归特征消除法。

    3）嵌入法（Embedded）

    嵌入法，先使用某些机器学习的算法和模型进行训练，得到各个特征的权重系数，根据系数从大到小选择特征；常用方法主要是基于惩罚项的特征选择法。

五、降维

    降维方法主要有：主成分析法(PCA)和线性判别分析法(LDA)。