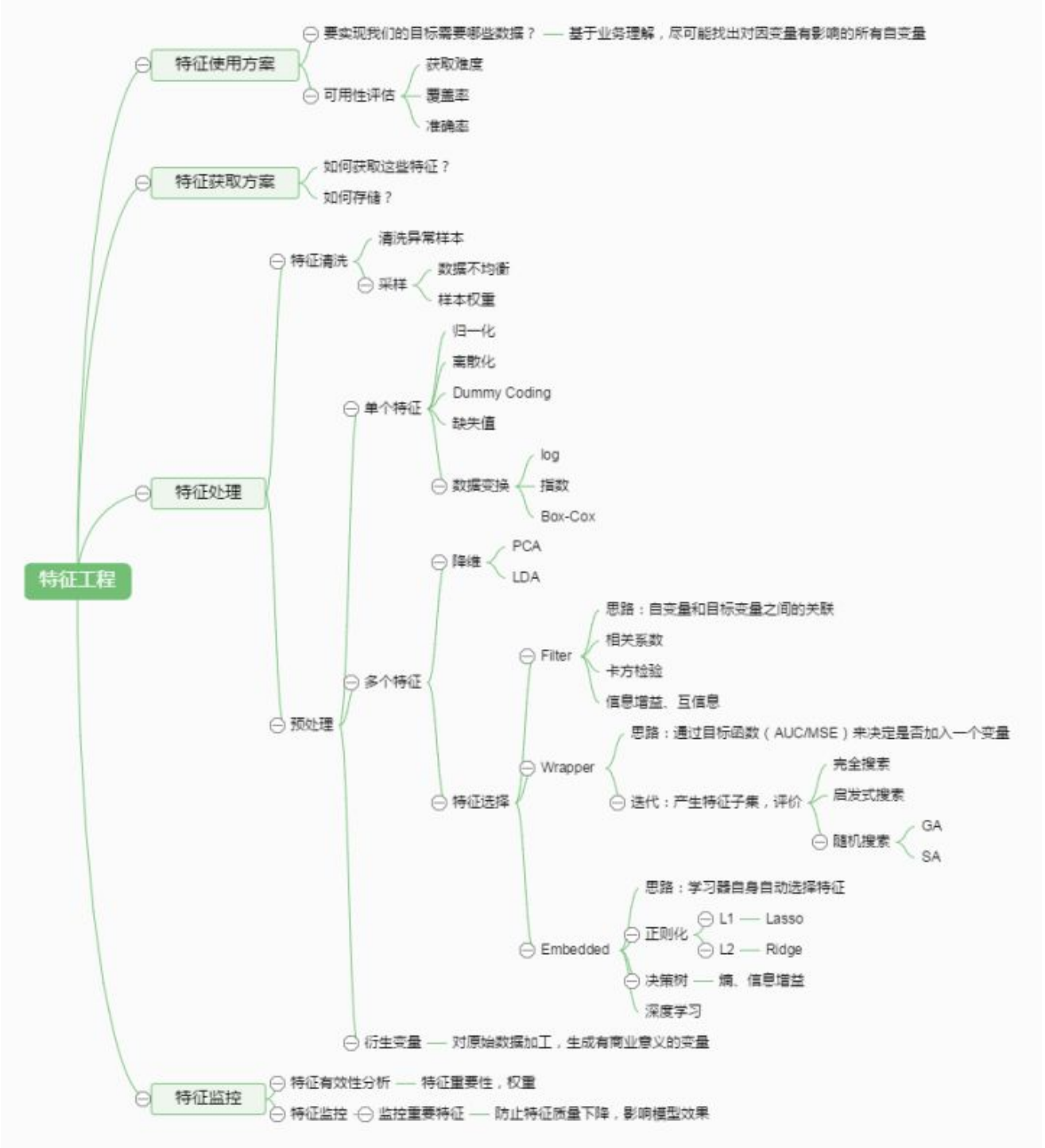
https://www.zhihu.com/question/28641663/answer/110165221

https://www.zhihu.com/question/28641663/answer/110165221

1.特征工程是什么?

构建特征工程的目的是最大限度地从原始数据中提取特征以供算法和模型使用。特征工程包括以下方面：



特征处理是特种工程的核心部分，sklearn提供了较为完整的特征处理方法。包括数据预处理，特征选择，降维等。

2.数据预处理

未处理的特征或数据，可能有以下问题：

1. 不属于同一量纲：特征的规格不一样，不能放在一起比较。无量纲化可以解决这个问题

2. 定性特征不能直接使用：定性转为定量，通常使用哑编码的方式将定性转为定量特征：假设有N种定性值，则将这一个特征扩展为N种特征，当原始特征值为第i种定性值时，第i个扩展特征赋值为1，其他扩展特征值为0.

3. 存在缺失值：需要补充

一般来说，sklearn中的preprocessing库可以解决以上问题

2.1 无量纲化

常用方法有标准化法和区间缩放法。标准化的前提是特征值服从正态分布，标准化后，其转换成标准正太分布。区间缩放法利用了边界值信息，将特征的取值区间缩放到某个特定范围。

2.1.1 标准化

```
from sklearn.preprocessing import StandardScaler

#标准化，返回值为标准化后的数据
StandardScaler().fit_transform(iris.data)
```

2.1.2 区间缩放法

$$x' = \frac{x - Min}{Max - Min}$$

```
from sklearn.preprocessing import MinMaxScaler

#区间缩放，返回值为缩放到[0, 1]区间的数据
MinMaxScaler().fit_transform(iris.data)
```

2.1.3 标准化与归一化的区别

简单来说，标准化是依照特征矩阵的列处理数据，其通过求z-score的方法，将样本的特征值转换到同一量纲下。归一化是依照特征矩阵的行处理数据，其目的在于样本向量在点乘运算或其他核函数计算相似性时，拥有统一的标准，也就是说都转化为“单位向量”。规则为l2的归一化公式如下：

$$x' = \frac{x}{\sqrt{\sum_j^m x[j]^2}}$$

```
from sklearn.preprocessing import Normalizer

#归一化，返回值为归一化后的数据
Normalizer().fit_transform(iris.data)
```

2.2 对定量特征二值化

$$x' = \begin{cases} 1, x > threshold \\ 0, x \leq threshold \end{cases}$$

```
from sklearn.preprocessing import Binarizer

#二值化，阈值设置为3，返回值为二值化后的数据
Binarizer(threshold=3).fit_transform(iris.data)
```

2.3 编码

```
from sklearn.preprocessing import OneHotEncoder
#哑编码，对IRIS数据集的目标值，返回值为哑编码后的数据
OneHotEncoder().fit_transform(iris.target.reshape((-1,1)))
```

2.4 缺失值

```
from numpy import vstack, array, nan
from sklearn.preprocessing import Imputer
#缺失值计算，返回值为计算缺失值后的数据
#参数missing_value为缺失值的表示形式，默认为NaN
#参数strategy为缺失值填充方式，默认为mean（均值）
Imputer().fit_transform(vstack((array([nan, nan, nan, nan]), iris.data)))
```

缺失数据一般有以下几种处理方法：

1. 删除包含空值的行和列，但一般会造成数据浪费
2. 自动补全，用平均值 最大值 最小值等填充
3. 手动补全，仅适用于数据量比较小的情况

3.特征选择

当数据预处理完成后，我们需要选择有意义的特征输入模型，通常从两个方面来考虑选择特征：

- 特征是否发散，如果一个特征不发散，例如方差接近0，也就是说样本在这个特征上基本没有差异，这个特征对于样本的区分没有什么用
- 特征与目标的相关性：这点比较显见，与目标相关性高的特征应当优先选择

根据特征选择的形式可以将特征选择方法分为3种：

- Filter：过滤法，按照发散性或者相关性对各个特征进行评分，设定阈值或者带选择阈值的个数，选择特征
- 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
- Embedder：嵌入法，先使用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，根据系数从大到小选择特征。类似于Filter方法，但是是通过训练来确定特征的优劣。

我们使用sklearn中的feature_selection来进行特征选择

3.1 Filter

3.1.1 方差选择法

使用方差选择法，先要计算各个特征的方差，然后根据阈值，选择方差大于阈值的特征。

```
rom sklearn.feature_selection import VarianceThreshold
#方差选择法，返回值为特征选择后的数据
#参数threshold为方差的阈值
VarianceThreshold(threshold=3).fit_transform(iris.data)
```

3.1.2 相关系数法

使用相关系数法，先要计算各个特征对目标值的相关系数以及相关系数的P值。

```
from sklearn.feature_selection import SelectKBest
from scipy.stats import pearsonr
#选择K个最好的特征，返回选择特征后的数据
#第一个参数为计算评估特征是否好的函数，该函数输入特征矩阵和目标向量，输出二元组（评分，P值）的数组，数组第i项为第i个特征的评分和P值。在此定义为计算相关系数
#参数k为选择的特征个数
SelectKBest(lambda X, Y: array(map(lambda x:pearsonr(x, Y), X.T)).T, k=2).fit_transform(iris.data, iris.target)
```

3.1.3 卡方检验

3.1.4 互信息法

3.2 Wrapper

3.2.1 递归特征消除法

使用一个基模型来进行多轮训练，每轮训练后，消除若干权值系数的特征，再基于新的特征集进行下一轮训练。

```
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
#递归特征消除法，返回特征选择后的数据
#参数estimator为基模型
#参数n_features_to_select为选择的特征个数
RFE(estimator=LogisticRegression(), n_features_to_select=2).fit_transform(iris.data, iris.target)
```

3.3 Embedded

3.3.1 基于惩罚项的特征选择法

使用带惩罚项的基模型，除了筛选出特征外，同时进行了降维。用feature_selection库的SelectFromModel类结合带L1惩罚项的逻辑回归模型，来选择特征的代码如下

```
from sklearn.feature_selection import SelectFromModel
 from sklearn.linear_model import LogisticRegression
 #带L1惩罚项的逻辑回归作为基模型的特征选择
 SelectFromModel(LogisticRegression(penalty="l1", C=0.1)).fit_transform(iris.data, iris.target)
```

3.3.2 基于树模型的特征选择法

树模型中GBDT也可用来作为基模型进行特征选择

```
from sklearn.feature_selection import SelectFromModel
from sklearn.ensemble import GradientBoostingClassifier
#GBDT作为基模型的特征选择 S
selectFromModel(GradientBoostingClassifier()).fit_transform(iris.data, iris.target)
```

4.降维

常见的降维方法除了以上提到的基于L1惩罚项的模型以外，另外还有主成分分析法（PCA）和线性判别分析（LDA），线性判别分析本身也是一个分类模型。PCA和LDA有很多的相似点，其本质是要将原始的样本映射到维度更低的样本空间中，但是PCA和LDA的映射目标不一样：PCA是为了让映射后的样本具有最大的发散性；而LDA是为了让映射后的样本有最好的分类性能。所以说PCA是一种无监督的降维方法，而LDA是一种有监督的降维方法。