



温州大學
WENZHOU UNIVERSITY

机器学习-特征工程

黄海广 副教授

2021年09月

本章目录

2

01 相关概念

02 特征构建

03 特征提取

04 特征选择

1. 相关概念

3

01 相关概念

02 特征构建

03 特征提取

04 特征选择

1. 相关概念

4

特征工程相关概念

定义

是把**原始数据**转变为模型的**训练数据**的过程

目的

获取更好的训练数据特征，使得机器学习模型逼近这个上限

作用

- 使模型的性能得到提升
- 在机器学习中占有非常重要的作用

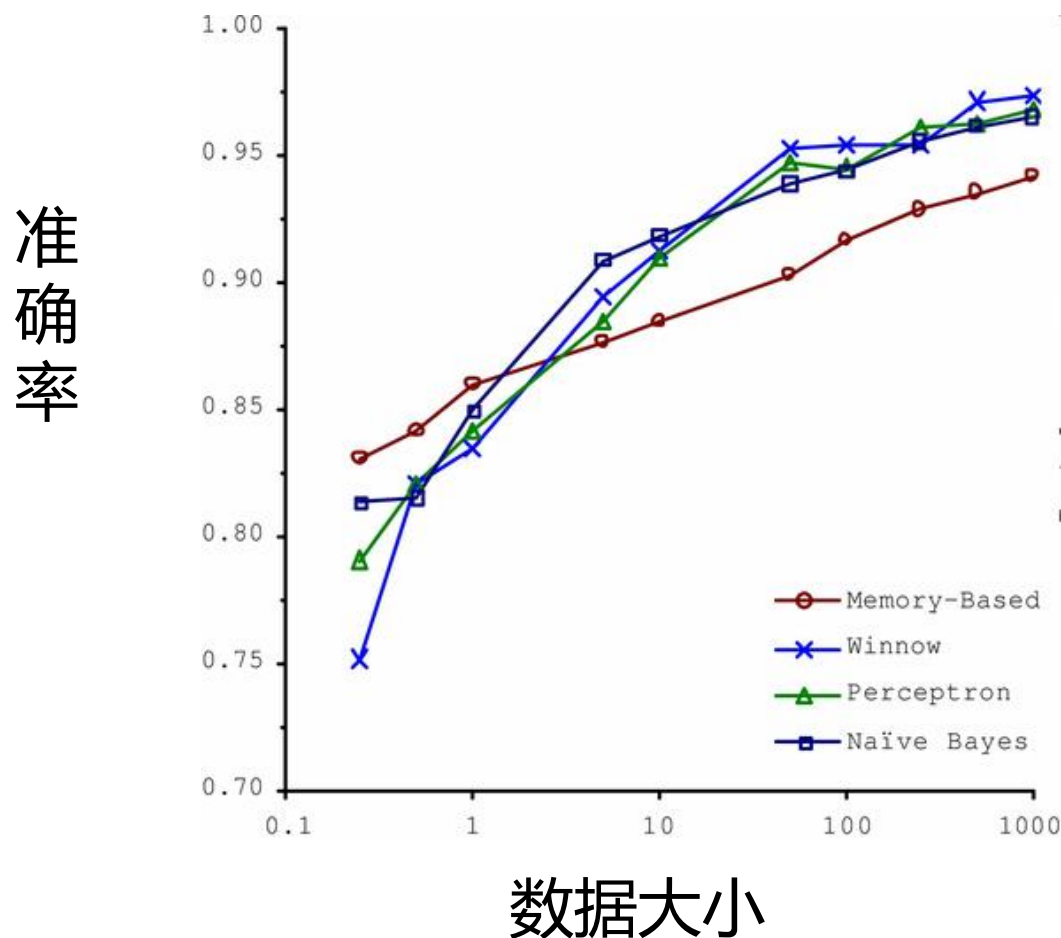
构成

- 特征构建
- 特征提取
- 特征选择

1. 相关概念

5

数据决定一切



通过这张图可以看出，各种不同算法在输入的数据量达到一定级数后，都有相近的高准确度。于是诞生了机器学习界的名言：

成功的机器学习应用不是拥有最好的算法，而是拥有最多的数据！

1. 相关概念

6

特征提取VS特征选择

项目	特征提取	特征选择
共同点	都从原始特征中找出最有效的特征 都能帮助减少特征的维度、数据冗余	
区别	<ul style="list-style-type: none">➤ 强调通过特征转换的方式得到一组具有明显物理或统计意义的特征➤ 有时能发现更有意义的特征属性	<ul style="list-style-type: none">➤ 从特征集合中挑选一组具有明显物理或统计意义的特征子集➤ 能表示出每个特征对于模型构建的重要性

2. 特征构建

7

01 相关概念

02 特征构建

03 特征提取

04 特征选择

2. 特征构建

8

在原始数据集中的特征的形式不适合直接进行建模时，使用一个或多个原特征构造新的特征可能会比直接使用原有特征更为有效。

特征构建：是指从原始数据中人工的找出一些具有物理意义的特征。

操作：使用混合属性或者组合属性来创建新的特征，或是分解或切分原有的特征来创建新的特征

方法：经验、属性分割和结合

2. 特征构建

9

数据规范化 使不同规格的数据转换到同一规格。

归一化（最大 - 最小规范化）

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

将数据映射到[0,1]区间

数据归一化的目的是使得各特征对目标变量的影响一致，会将特征数据进行伸缩变化，所以数据归一化是会**改变特征数据分布**的。

Z-Score标准化

$$x^* = \frac{x - \mu}{\sigma}$$

处理后的数据均值为0，方差为1

数据标准化为了不同特征之间具备可比性，经过标准化变换之后的**特征数据分布没有发生改变**。

就是当数据特征取值范围或单位差异较大时，最好是做一下标准化处理。

$$\sigma^2 = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)^2$$
$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$$

2. 特征构建

10

定量特征二值化

设定一个阈值，大于阈值的赋值为1，小于等于阈值的赋值为0，使用preprocessing库的Binarizer类对数据进行二值化的代码如下：

```
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. 特征构建

11

定性特征哑编码

使用preprocessing库的OneHotEncoder类对数据进行哑编码的代码如下：

```
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. 特征构建

12

分箱

一般在建立分类模型时，需要对连续变量离散化，特征离散化后，模型会更稳定，降低了模型过拟合的风险。

设成绩为：[63 64 88 71 42 60 99 70 32 88 34 69 83 52 66 92 82 58 66 41]

可以按照区间分箱：

```
bins=[0,59,70,80,90,100]
score_cat = pd.cut(score_list,
bins)
print(pd.value_counts(score_cat))
```

```
(59, 70] 7
(0, 59] 6
(80, 90] 4
(90, 100] 2
(70, 80] 1
```

也可以数量分箱：

```
score_cat = pd.qcut(score_list,5)
print(pd.value_counts(score_cat))
```

```
(31.999, 50.0] 4
(50.0, 63.6] 4
(63.6, 69.4] 4
(69.4, 84.0] 4
(84.0, 99.0] 4
```

2. 特征构建

13

聚合特征构造

- 聚合特征构造主要通过对多个特征的分组聚合实现，这些特征通常来自同一张表或者多张表的联立。
- 聚合特征构造使用一对多的关联来对观测值分组，然后计算统计量。
- 常见的分组统计量有中位数、算术平均数、众数、最小值、最大值、标准差、方差和频数等。

2. 特征构建

14

转换特征构造

相对于聚合特征构造依赖于多个特征的分组统计，通常依赖于对于特征本身的变换。转换特征构造使用单一特征或多个特征进行变换后的结果作为新的特征。

常见的转换方法有单调转换（幂变换、log变换、绝对值等）、线性组合、多项式组合、比例、排名编码和异或值等。

2. 特征构建

15

转换特征构造

此外，由于业务的需求，一些指标特征也需要基于业务理解进行特征构造。

- 基于单价和销售量计算销售额.
- 基于原价和售价计算利润.
- 基于不同月份的销售额计算环比或同比销售额增长/下降率.
-

3. 特征提取

16

01 相关概念

02 特征构建

03 特征提取

04 特征选择

3. 特征提取

17

提取对象：原始数据（特征提取一般是在特征选择之前）

提取目的：自动地构建新的特征，将原始数据转换为一组具有明显物理意义（比如几何特征、纹理特征）或者统计意义的特征。

常用方法

降维方面的PCA、ICA、LDA等

图像方面的SIFT、Gabor、HOG等

文本方面的词袋模型、词嵌入模型等

3. 特征提取

18

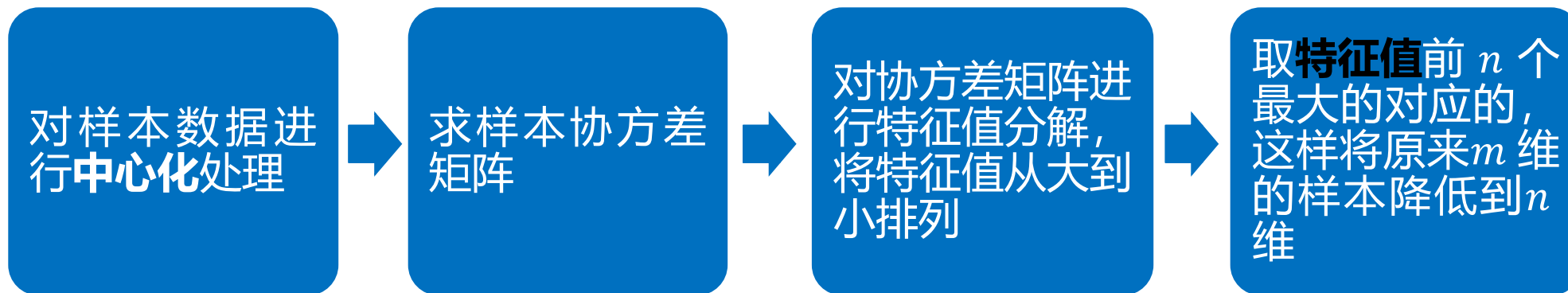
降维

1.PCA(Principal Component Analysis, 主成分分析)

PCA 是降维最经典的方法，它旨在找到数据中的主成分，并利用这些主成分来表征原始数据，从而达到降维的目的。

PCA 的思想是通过坐标轴转换，寻找数据分布的最优子空间。

步骤



3. 特征提取

19

降维

2. ICA(Independent Component Analysis, 独立成分分析)

ICA独立成分分析，获得的是相互独立的属性。ICA算法本质寻找一个线性变换 $z = Wx$ ，使得 z 的各个特征分量之间的**独立性最大**。

步骤

PCA 对数据
进行降维



ICA 来从多
个维度分离
出有用数据

PCA 是 ICA 的数据预处理方法

3. 特征提取

20

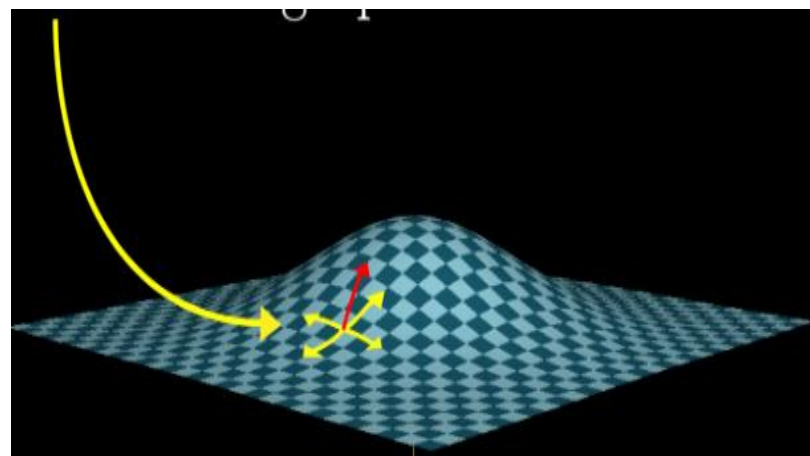
图像特征提取

1. SIFT 特征

优点:

- 具有旋转、尺度、平移、视角及亮度不变性，有利于对目标特征信息进行有效表达；
- SIFT 特征对参数调整鲁棒性好，可以根据场景需要调整适宜的特征点数量进行特征描述，以便进行特征分析。

缺点： 不借助硬件加速或者专门的图像处理器很难实现。



步骤

疑似特征点检测

去除伪特征点

特征点梯度
与方向匹配

特征描述向量的
生成

3. 特征提取

21

图像特征提取

2. HOG特征

方向梯度直方图(HOG)特征是 2005 年针对**行人检测问题**提出的直方图特征，它通过计算和统计图像**局部区域的梯度方向**直方图来实现特征描述。

步骤

归一化处理

计算图像梯度

统计梯度方向

特征向量
归一化

生成特征向量

3. 特征提取

22

文本特征提取

1. 词袋模型

将整段文本**以词为单位**切分开，然后每篇文章可以表示成一个长向量，向量的每一个维度代表一个单词，而该维度的权重反映了该单词在原来文章中的重要程度

采用 TF-IDF 计算权重，公式为 $TF - IDF(t, d) = TF(t, d) \times IDF(t)$

$TF(t, d)$ 表示单词 t 在文档 d 中出现的频率

$IDF(t)$ 是逆文档频率，用来衡量单词 t 对表达语义所起的重要性，其表示为：

$$IDF(t) = \log \frac{\text{文章总数}}{\text{包含单词}t\text{的文章总数} + 1}$$

3. 特征提取

23

文本特征提取

2. N-gram 模型

- 将连续出现的 n 个词 ($n \leq N$) 组成的词组(N-gram)作为一个单独的特征放到向量表示, 构成了 N-gram 模型。
- 另外, 同一个词可能会有多种词性变化, 但却具有相同含义, 所以实际应用中还会对单词进行词干抽取(Word Stemming)处理, 即将不同词性的单词统一为同一词干的形式。

4. 特征选择

24

01 相关概念

02 特征构建

03 特征提取

04 特征选择

4. 特征选择

25

特征选择(feature selection): 从给定的特征集合中选出相关特征子集的过程。

原因: 维数灾难问题; 去除无关特征可以降低学习任务的难度, 简化模型, 降低计算复杂度

目的: 确保不丢失重要的特征

相关特征

- 对当前学习任务有用的属性或者特征

无关特征

- 对当前学习任务没用的属性或者特征

4. 特征选择

26

模型性能

- 保留尽可能多的特征，模型的性能会提升
- 但同时模型就变复杂，计算复杂度也同样提升

VS

计算复杂度

- 剔除尽可能多的特征，模型的性能会有所下降
- 但模型就变简单，也就降低计算复杂度

4. 特征选择

27

特征选择的三种方法

过滤式(Filter):

先对数据集进行特征选择，其过程与后续学习器无关，即设计一些统计量来过滤特征，并不考虑后续学习器问题

包裹式(Wrapper):

就是一个分类器，它是将后续的学习器的性能作为特征子集的评价标准

嵌入式(Embedding):

是学习器自主选择特征

4. 特征选择

28

过滤式

原理：先对数据集进行特征选择，然后再训练学习器

特征选择过程与后续学习器无关

也就是先采用特征选择对初始特征进行过滤，然后用过滤后的特征训练模型

优点：计算时间上比较高效，而且**对过拟合问题**有较高的鲁棒性

缺点：倾向于选择冗余特征，即没有考虑到特征之间的相关性

4. 特征选择

29

过滤式

1、Relief 方法



- ◆ 定义：Relevant Features是一种著名的过滤式特征选择方法。该方法设计了一个相关统计量来度量特征的重要性。
 - 该统计量是一个向量，其中每个分量都对应于一个初始特征。
 - 特征子集的重要性则是由该子集中每个特征所对应的相关统计量分量之和来决定的。
 - 最终只需要指定一个阈值 k ，然后选择比 k 大的相关统计量分量所对应的特征即可。也可以指定特征个数 m ，然后选择相关统计量分量最大的 m 个特征。
- ◆ Relief 是为二分类问题设计的，其拓展变体 Relief-F 可以处理多分类问题。

4. 特征选择

30

过滤式

2、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。

3、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。

4、卡方检验



检验定性自变量对定性因变量的相关性。假设自变量有 N 种取值，因变量有 M 种取值，考虑自变量等于 i 且因变量等于 j 的样本频数的观察值与期望的差距，构建统计量：

$$X^2 = \sum \frac{(A - E)^2}{E}$$

4. 特征选择

31

过滤式

5、互信息法

概念：经典的互信息也是评价定性自变量对定性因变量的**相关性的**。

为了处理定量数据，最大信息系数法被提出。

互信息计算公式如下：

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

4. 特征选择

32

包裹式



原理：包裹式特征选择**直接把最终将要使用的学习器的性能作为特征子集的评价原则**。其目的就是为给定学习器选择最有利于其性能、量身定做的特征子集。



优点：直接针对特定学习器进行优化，考虑到特征之间的关联性，因此通常包裹式特征选择比过滤式特征选择能训练得到一个更好性能的学习器。

缺点：由于特征选择过程需要多次训练学习器，故计算开销要比过滤式特征选择要大得多。

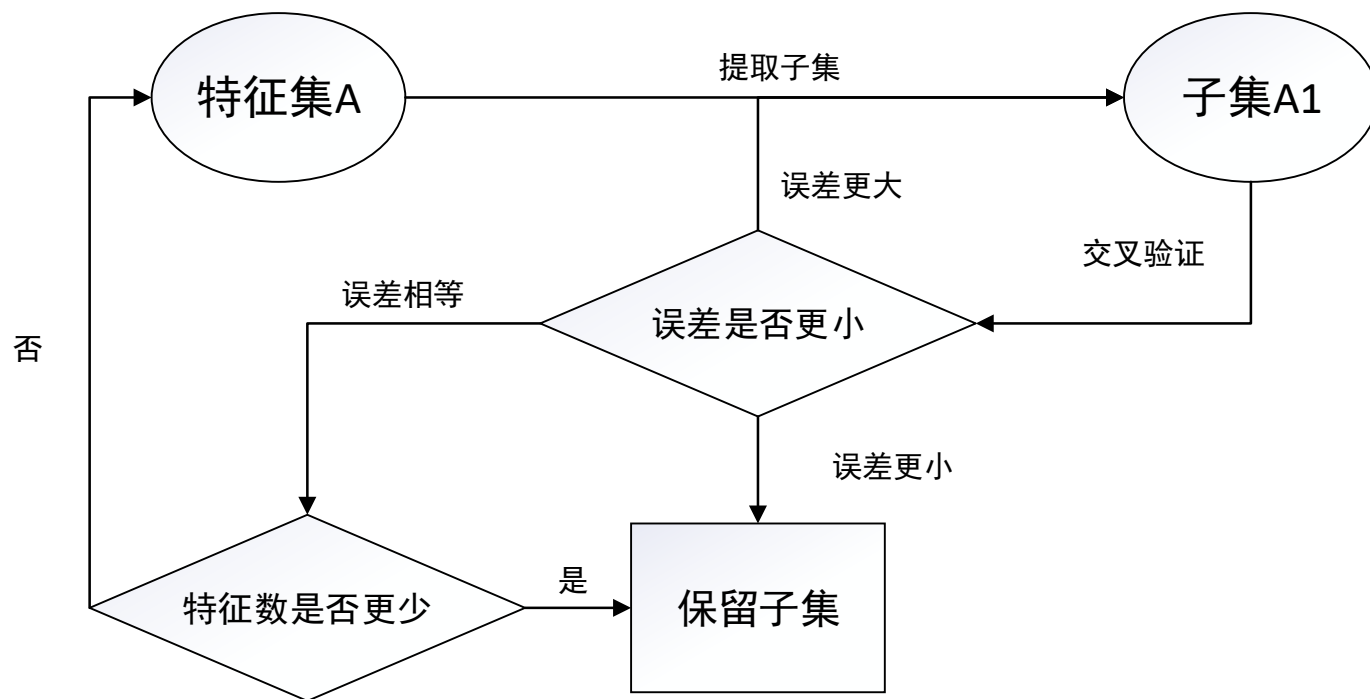
4. 特征选择

33

包裹式

1. LVW

- Las Vegas Wrapper是一个典型的包裹式特征选择方法。使用随机策略来进行子集搜索，并以**最终分类器的误差**作为特征子集的评价标准。
- 由于 LVW 算法中每次特征子集评价都需要训练学习器，计算开销很大，因此它会设计一个停止条件控制参数 T 。但是如果初始特征数量很多、 T 设置较大、以及每一轮训练的时间较长，则很可能算法运行很长时间都不会停止。



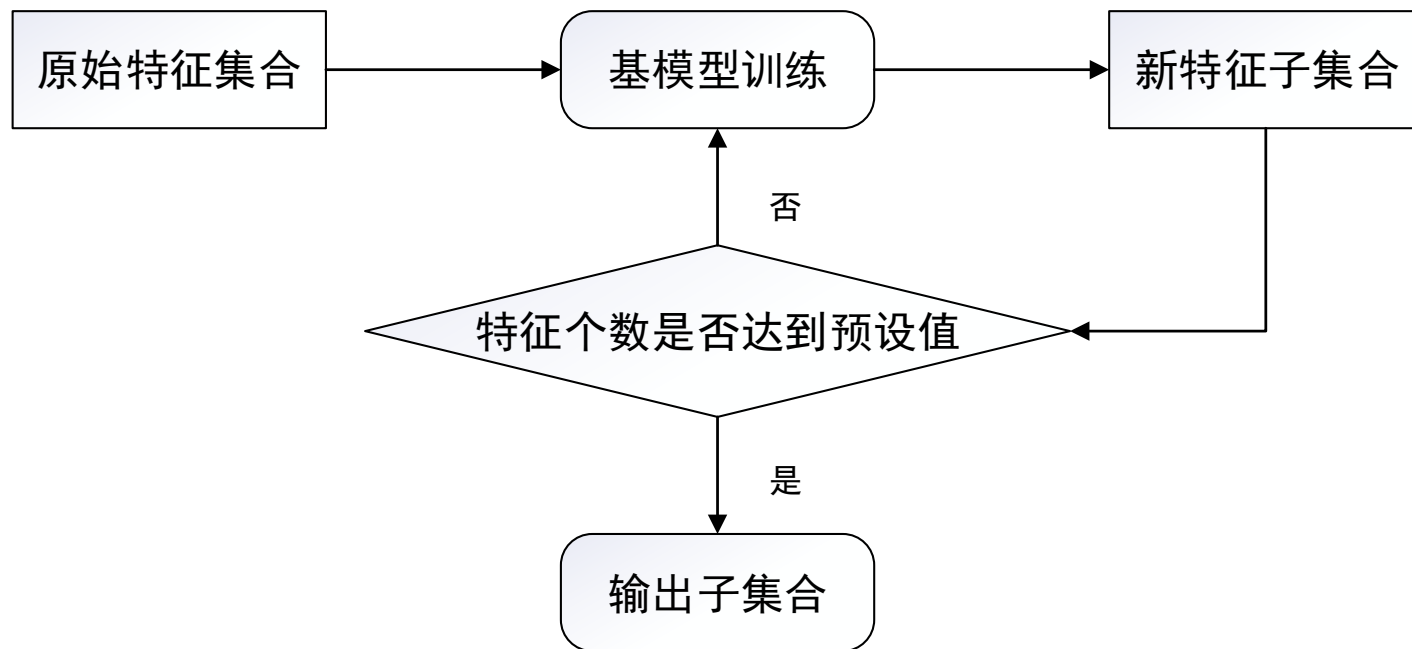
4. 特征选择

34

包裹式

2. 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。



4. 特征选择

35

嵌入式



原理：嵌入式特征选择是将特征选择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的。即学习器训练过程中自动进行了特征选择。



常用的方法包括：

- 利用**正则化**，如L1, L2 范数，主要应用于如线性回归、逻辑回归以及支持向量机(SVM)等算法；优点：降低过拟合风险；求得的 w 会有较多的分量为零，即：它更容易获得稀疏解。
- 使用决策树思想，包括决策树、随机森林、Gradient Boosting 等。

4. 特征选择

36

嵌入式

常见的嵌入式选择模型：



在 Lasso 中， λ 参数控制了稀疏性：

- 如果 λ 越小，则稀疏性越小，被选择的特征越多
- 相反 λ 越大，则稀疏性越大，被选择的特征越少



在 SVM 和 逻辑回归中，参数 C 控制了稀疏性：

- 如果 C 越小，则稀疏性越大，被选择的特征越少
- 如果 C 越大，则稀疏性越小，被选择的特征越多

1. Prof. Andrew Ng. Machine Learning. Stanford University
2. 《统计学习方法》，清华大学出版社，李航著，2019年出版
3. 《机器学习》，清华大学出版社，周志华著，2016年出版
4. 《特征工程及 XGBoost模型》，武汉理工大学课件

谢谢!

