人工智能之机器学习

特征工程

主讲人：李老师

学习目标

- 了解特征工程在机器学习当中的重要性
- 特征预处理
- 特征提取
- 特征选择和特征的降维

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

特征工程相关概念

- 为什么要做特征工程？？？
  - 数据决定一切：机器学习领域的大神Andrew Ng(吴恩达)老师说

    “Coming up with features is difficult, time-consuming, requires expert knowledge. ‘Applied machine learning’ is basically feature engineering.”
  - 业界广泛流传：数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。

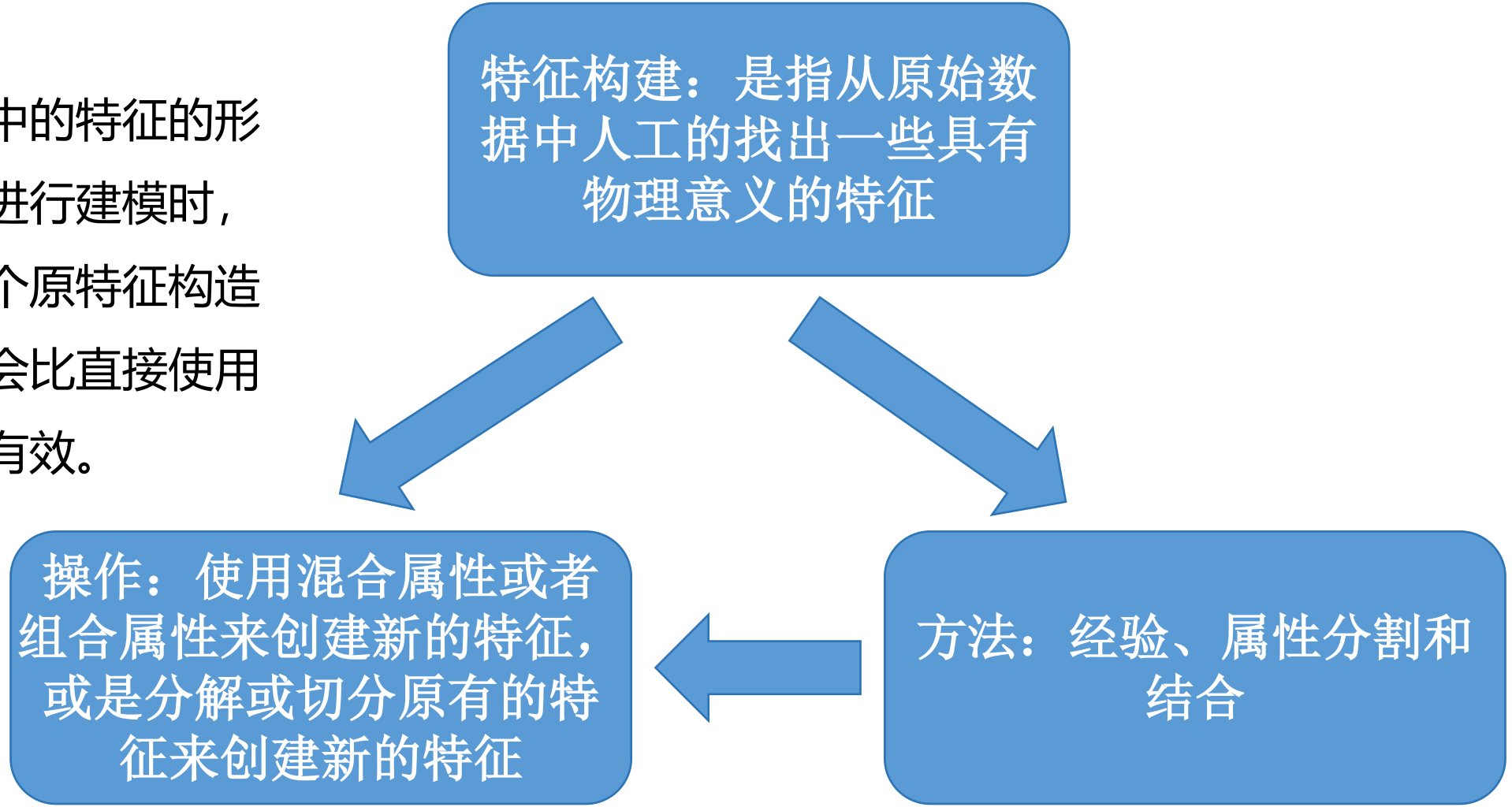
特征工程相关概念

- 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	➢ 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	➢ 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

特征构建

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

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

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

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

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

处理后的数据均值为0，方差为1

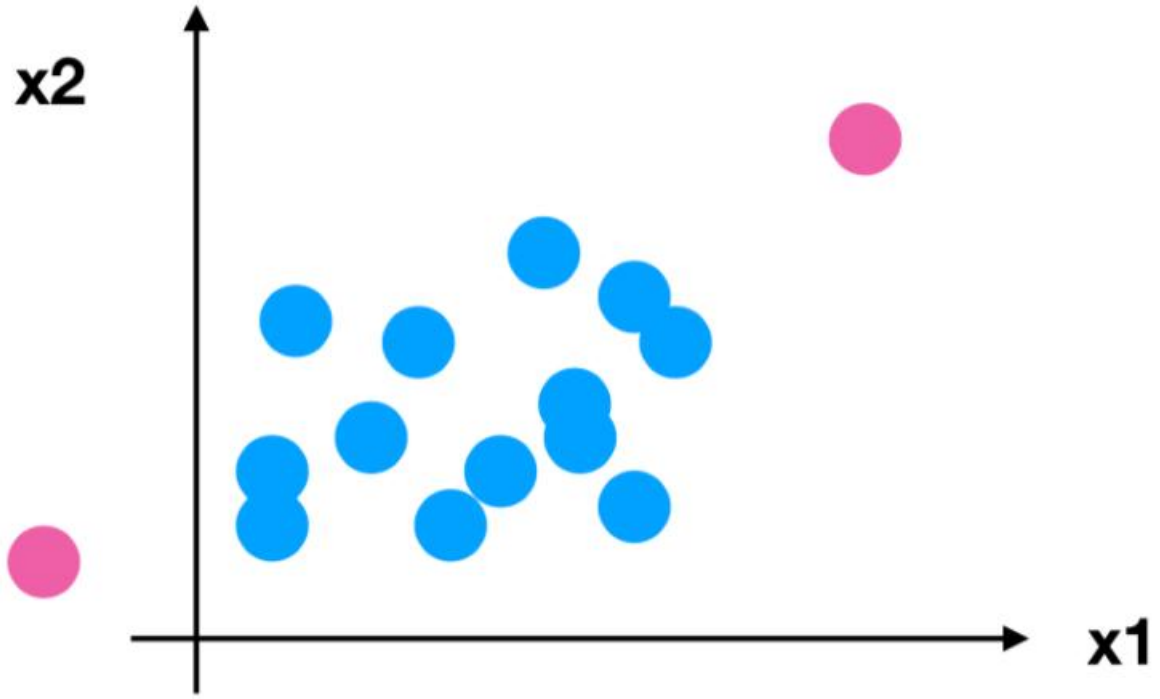
数据标准化为了不同特征之间具备可比性，经过标准化变换之后的特征数据分布没有发生改变。
就是当数据特征取值范围或单位差异较大时，最好是做一下标准化处理。

归一化和标准化

- 为什么我们要进行归一化/标准化?
  - 特征的单位或者大小相差较大，或者某特征的方差相比其他的特征要大出几个数量级，容易影响（支配）目标结果，使得一些算法无法学习到其它的特征
  - 我们需要用到一些方法进行无量纲化，使不同规格的数据转换到同一规格

归一化

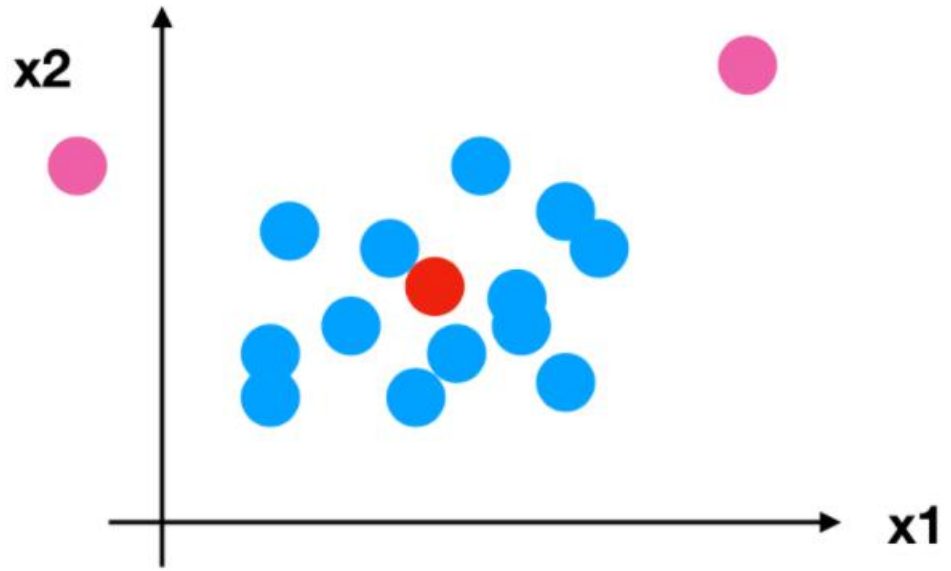
- 思考：如果数据中异常点较多，会有什么影响？



注意最大值最小值是变化的，另外，最大值与最小值非常容易受异常点影响，所以这种方法鲁棒性较差，只适合传统精确小数据场景。

标准化

- 回到刚才异常点的地方，我们再来看看标准化



对于归一化来说：如果出现异常点，
影响了最大值和最小值，那么结果显然会发生改变

对于标准化来说：如果出现异常点，
由于具有一定数据量，
少量的异常点对于平均值的影响并不大，
从而方差改变较小。

- 在已有样本足够多的情况下比较稳定，适合现代嘈杂大数据场景。

特征构建

- 特征二值化：连续变量离散化
- 分箱：连续变量离散化
- 特征哑编码：离散特征的处理

特征构建

- 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。

特征构建

- **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。
  - 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.
    - ……

特征提取

- 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。
- 例：
  - 图像方面的SIFT、Gabor、HOG等
  - 文本方面的词袋模型、词嵌入模型等

文本特征提取

- 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
- TF-IDF
  - $TF-IDF(t, d) = TF(t, d) \times IDF(t)$
  - $TF(t, d)$ 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：

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

特征选择&降维

- **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信息。
- **降维**：
  - PCA 是降维最经典的方法，它旨在是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。

相关特征	无关特征
• 对当前学习任务有用的属性或者特征	• 对当前学习任务没用的属性或者特征

模型性能	VS	计算复杂度
• 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		• 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

特征选择

- 特征选择的方法主要有以下三种：
  - 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。
  - 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。
  - 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。

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

特征选择-过滤法

- 方差选择法：先计算各个特征属性的方差值，然后根据阈值，获取方差大于阈值的特征。
- 相关系数法：先计算各个特征属性对于目标值的相关系数以及阈值K，然后获取K个相关系数最大的特征属性。(备注：根据目标属性y的类别选择不同的方式)
- 卡方检验：检查定性自变量对定性因变量的相关性。

特征选择-包装法

- 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。

特征选择-嵌入法

- 使用基于惩罚项的基模型，进行特征选择操作。

特征降维-PCA

- PCA(Principal Component Analysis)是常用的线性降维方法，是一种无监督的降维算法。算法目标是通过某种线性投影，将高维的数据映射到低维的空间中表示，并且**期望在所投影的维度上数据的方差最大（最大方差理论）**，以此使用较少的数据维度，同时保留较多的原数据点的特性。
- 通俗来讲的话，如果将所有点映射到一起，那么维度一定降低下去了，但是同时也会将几乎所有的信息(包括点点之间的距离等)都丢失了，而如果映射之后的数据具有比较大的方差，那么可以认为数据点则会比较分散，这样的话，就可以保留更多的信息。从而我们可以看到PCA是一种丢失原始数据信息最少的无监督线性降维方式。

特征降维-PCA

- 在PCA降维中，数据从原来的坐标系转换为新的坐标系，新坐标系的选择由数据本身的特性决定。第一个坐标轴选择原始数据中方差最大的方向，从统计角度来讲，这个方向是最重要的方向；第二个坐标轴选择和第一个坐标轴垂直或者正交的方向；第三个坐标轴选择和第一个、第二个坐标轴都垂直或者正交的方向；该过程一直重复，直到新坐标系的维度和原始坐标系维度数目一致的时候结束计算。而这些方向所表示的数据特征就被称为“主成分”。

特征降维-PCA

- 主成分析(PCA)：将高纬的特征向量合并称为低纬度的特征属性，是一种无监督的降维方法。