**特征工程**

特征工程是指从原始数据转换为特征向量的过程。特征工程是机器学习中最重要的起始步骤，会直接影响机器学习的效果，并通常需要大量的时间。典型的特征工程包括数据清理、特征提取、特征选择等过程。

特征特点：1有固定不变的，如商品颜色，品牌，

2有随时间变化的，如销量，排名，时间变化采样的颗粒度要按照实际推荐效果来决定，很可能过去三个月的销售排名对推荐效果来说可以很稳定，也或许昨天的排名对今天的推荐效果比三个月平均更好。

3有周期性的

4直接特征，

5间接特征，如单反相机的购买记录里，男性的比例会高于女性，所以单反相机的性别特征向量可以是购买人数性别的比例值。

**数据清理**

* 缩放特征值（归一化）：将浮点特征值从自然范围（如 100 到 900）转换为标准范围（如 0 到 1）。特征集包含多个特征时，缩放特征可以加快梯度下降的收敛过程，并可以避免 NaN 陷阱。特征缩放的方法一般为 scaled-value = (value-mean)/stddev*scaled*−*value*=(*value*−*mean*)/*stddev*
* 处理极端离群值，如取对数、限制最大最小值等方法
* 分箱（离散化）
* 填补遗漏值
* 移除重复样本、不良标签、不良特征值等
* 平滑
* **特征提取：**

1. 理解数据
   1. 特征是连续的还是离散
   2. 特征数据的分布如何
   3. 分布依赖的因素
   4. 是否有数据缺失，重复，交叉
   5. 特征的来源
   6. 数据是实时的吗，（固定，实时，周期，）
2. 头脑风暴更多的特征  
   好的特征具有下面特点
   1. 能够直观地解释
   2. 可以被计算
   3. 是很好地观察数据的方式  
      例如：用户是否成为网站的注册用户？过去一周花费多少时间在这个网站上？在所有注册用户花费时间的分布中，这个用户处于什么位置？

* 5间接特征，如单反相机的购买记录里，男性的比例会高于女性，所以单反相机的性别特征向量可以是购买人数性别的比例值。
* 6组合特征（交叉特征）

**特征选择：**

* 从特征集合中挑选一组最具统计意义的特征子集，把无关的特征删掉，从而达到降维的效果
* 常用的方法：
* filter（刷选器）方法：Pearson相关系数，Gini-index（基尼指数），IG（信息增益）等
* wrapper（封装器）：向前选择（Forward selection）和向后选择（Backward selection）等
* Embeded(集成方法)：Regularization，或者使用决策树思想，Random Forest和Gradient boosting等
* 基于L1，L2惩罚项的特征选择法，
* 降维：主成分分析法（PCA），线性判别分析法（LDA）