数据清洗

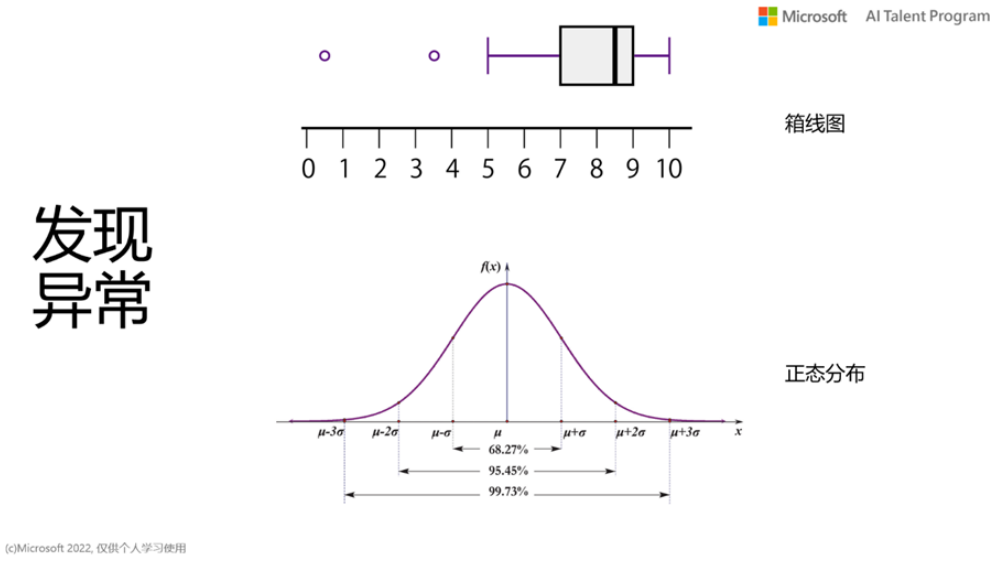
该课程主要为大家讲授如下的内容：

- 冗余值
- 异常处理
- 缺失值

1. 冗余值
    1. 冗余值
       冗余值就是数据多余的、重复的值。
       冗余表现为完全冗余和部分冗余。完全冗余指两个条数据一模一样，而部分冗余指其中某两条数据的某些字段值一样；从另一个角度看，可以分为样本冗余和特征的冗余。在数据清洗模块，只考虑样本的冗余，特征的冗余需要在特征工程时对特征进行筛选。

       通常情况下，冗余数据是由于不同来源数据表的合并造成的。冗余值不但会造成计算、存储的压力增大，还可能隐含数据不一致的问题。
    2. 处理方法
       对于冗余值，最简单的处理方式就是将冗余的样本直接去除。
2. 异常处理
    1. 异常值和错误值
       一般来说，异常处理涉及异常值和错误值。
       错误值指不符合数据原始假设的值，是异常值的一种；而异常值通常指不寻常的数据，比如距离数据分布较偏远的数据点，这一类可以被称为“外点（outlier）”。
       由于异常数据并不总是错误值，因此我们还可以建立一个用于异常数据处理的小型pipeline，先对数据设定一个区间（或是一些其他的规则），如果超出区间的进行异常处理，判定是否为错误数据，再进行接下来的处理。

2. 发现异常
   如何发现异常值在机器学习中被统称为异常检测问题。
   识别异常数据的方法：
   - 绘制箱线图（box plot）；
   - 根据中心极限定理，将数据转化为正态分布，若样本点落在距离均值3个sigma（标准差）的距离的区域，此时的可能性小于0.3%，就可被判定为异常点。



3. 处理异常
   处理异常的方法：
   - 直接将异常值去除，作为缺失值对待。这种方法在可视化分析中会经常用到。
   - 保留异常值。如果异常值不是由于数据收集而产生错误，而是存在更深层次的原因导致异常值的出现，那么异常值将包含极大的信息量。通过收集更多信息分析异常值的出现，可能会得到非常重要而有趣的结论。因此，适当保留异常值也可能成为数据分析的关键。
   - 当异常值被判定为错误值时直接去除。

3. 缺失值
   1. 缺失值
      缺失值产生于数据的人工录入、机器故障、传输错误等诸多因素。
      缺失值的表现形式并不只限于数据项为空。
   2. 处理方法

1. 去除
    - 去除样本
    - 去除特征
2. 填充
    - 设定default值进行填充
    - 向前/向后填充
    - 序列数据，插值填充
    - EM算法，机器学习模型预测
