数据预处理

数据预处理：数据预处理是将原始数据转换为更适合模型输入的数据形势的过程，主要目的是清洗，数据规范化和转换数据，确保模型可以**正确处理**并从中提取信息。

常见步骤：

1） 处理缺失值（填补，删除）

2） 处理异常值（裁剪，平滑）

3） 数据格式转换（日期转时间戳，分类变量编码）

4） 数据缩放（标准化，归一化）

5） 数据集划分（训练集，验证集，测试集）

特征工程：对数据进行深入处理以构造更有意义的特征或提取新的信息，主要是**提升模型性能**，使模型能够更好地捕捉数据中的模式。

常见步骤：

1） 特征提取：从原始数据中生成新的特征（文本嵌入，图像特征）

2） 特征选择：剔除冗余，不相关或高度相关的特征

3） 特征转换：例如通过主成分分析（PCA）降维

4） 构造特征：结合已有特征创建新的特征（如时间差，比值，交互项）

그럼 전처리와 특징공정의 관계는 무엇일까?

层次关系：数据预处理是数据准备的基础部分，通常在特征工程之前完成。特征工程在数据清理后进行，对清洗后的数据进行更深入的处理和优化。

目标侧重：

-数据预处理：让数据达到可用的状态，解决**基础数据问题**。

-特征工程：让特征变得更有效，挖掘数据的潜在价值，**提升模型性能**。

流程上的协同：

-数据预处理：可以为特征工程提供高质量的输入数据

-特征工程：可能涉及对数据进一步预处理（如：特征缩放，编码）

왜 전처리를 해야 하는가?

데이터가 많아지면 데이터가 不完整，含噪声，不一致하는 문제가 생길수 있음, 그래서 질 좋은 데이터가 없으면 질 좋은 挖掘결과가 없음(즉, 提高数据挖掘的质量，精度)하기 위해서

常用方法:

1) 数据清理：填写缺失值，平滑噪声数据，识别，删除孤立点，解决不一致性
2) 数据集成：集成多个数据库，数据立方体或文件
3) 数据变换：将数据转换或统一成适合于挖掘的形式。如数据规范化
4) 数据归约：可以用来得到数据集的归约（压缩）表示，虽然小，但保持数据完整性。对归约后的数据集挖掘将更有效，并产生相似的分析结构。
5) 数据离散化于概念分层：数据归约的一部分，通过数据的离散化和概念分层来归约数据

数据清理

1. 空缺值처리 방법

-忽略元组：걍 빼버리고 하는 경우, 보통 이렇게 하지만 결측치 비율이 클 때 효과가 안좋을 수 있음

-人工填充空缺值：작업량 많아서, 실행하기 어려움

-使用一个全局变量填充：如unknown, -infinity이런걸로

-用平均值填充：

-用于给定元组同一类的所有样本的平均值

-使用最可能的值田中空缺值：使用Bayesian公式或判定树等基于推断方法

2. 噪声처리 방법(一个测量变量中的随机错误或偏差)

-**分箱（binning）：**첨에는 데이터 배열하고 分到等深的箱中，그 후에 按箱平均值平滑，按箱中值平滑，按箱的边界平滑等等

-**聚类**：通过聚类分析查找孤立点，并去除

-计算机和人工检查结合：计算机检测可疑数据，然后对它们进行人工判断

-**回归**：通过让数据适应回归函数来平滑数据（그니까 함수로 데이터를 표현하는거）

数据集成

步骤一， 数据集成：将多个数据源中的苏剧整合到一个一致的存储中

步骤二， 模式集成：整合不同数据源中的元数据

步骤三， 检测并解决数据值的冲突：解决不同的数据表示，不同的度量

이거 하다 보면 필연적으로 冗余数据가 나오게 되어있는데 어떻게 처리하지?

일단 **相关分析**로 부분적 冗余数据를 찾을수 있음

그리고 꼼꼼하게 고려해서 集成하면 데이터의 불일치나 冗余를 방지할수 있어

数据变换

数据变换：将数据转换或统一成适合于挖掘的形式

规范化：将数据按比例缩放，使之落入一个小的特定区间

1) Min-max规范化（如【0， 1】区间）

2) Z-score规范化（正态分布，Z（0，1））

数据归约

数据归约：可以用来得到数据集的归约表示，它小得多，但可以产生相同的（分析结果）

数据归约策略：

1) 数据立方体聚集

2) 维归约

3) 数据压缩

4) 数值归约

5) 直方图

6) 聚类

7) 分层选样

离散化和概念分层

离散化：通过将属性域划分为区间，减少给定连续属性值的个数。区间得标号可以代替实际的数据值。

概念分层：通过使用高层的概念来替代底层的属性值来规约数据（나이를 청소년, 중년, 노인 으로）

3个属性值类型：名称型，序数（순서가있는 명사, 직위같은），连续值(实数)

数据数值得离散化：

1) 分箱

2) 直方图分析

3) 聚类分析

4) 基于熵得离散化

5) 通过自然划分分段

通过自然划分分段：将数值区域划分为相对一致的，易于阅读的，看上去更直观或自然的区间

3-4-5规则：用于将数值数据划分为相对一致和“自然的”区间

步骤：

1） 如果一个区间最高有效位上包含 3，6，7，9个不同的值， 划分为**3**个等宽子区间（7은 2,3,2）

2） 如果2，4，8 → **4**

3） 1，5，10 → **5**

4） 将该规则递归地应用于每个子空间，产生给定数值属性的概念分层。

次序数据的定量化方法（시험에 나왔음）

次序指标：反应研究对象的所有个体在某一性质或特征基础下的具体表现程度的离散顺序的代表值

次序数据的定量化方法是指**将指数进行排序后得到的序号转化为正向的数量型数据的方法**

对于次序评价指标的定量化，可按下述公式转换

$$y_i = 1 - \frac{1}{n}(x_i - 0.5) \qquad (1.3.1)$$

式中 x_i —— 被评价对象的名次

y_i —— 第 个被评价对象的评价得分；

n —— 所有参评对象的个数。

等级指标→ **系数转换法**

在处理等级评价指标时，对于各个等级的划分使用对应的系数转换为无量纲的定量数据的方法。