Association Rule 关键定义

- **Itemset**：一组一个或多个 items（项）的集合。
  - **k-itemset**：包含 k 个 items 的 itemset，例如 X = {x_1, ..., x_k}。
- **Support**（支持度）：
  - **Absolute Support (Count)**：包含该 itemset 的 transaction 数量。
  - **Relative Support**：包含该 itemset 的 transaction 占总 transaction 的比例。
    - 记作 $\mathrm{sup}(\mathrm{X})$，即 transaction 包含 itemset X 的概率。
    - $\mathrm{sup}(\mathrm{X}) = \frac{\mathrm{count}(\mathrm{X})}{\mathrm{count}(\mathrm{Total})}$
      - $\mathrm{sup}(\mathrm{X})$ 频率
      - $\mathrm{P}(\mathrm{X})$ 概率
- **Frequent Itemset**：
  - 如果 itemset X 的 support 不小于最小支持度阈值（minsup），则称 X 为 frequent itemset。
- **Association Rule**（关联规则）：
  - 形式为 $\mathrm{X} \rightarrow \mathrm{Y}$，其中 X 和 Y 都是非空 itemset，且 $\mathrm{X} \cup \mathrm{Y} = \emptyset$。
  - 描述了“如果 X 出现，则 Y 也很可能出现”的关系。
- **Support of Rule**（规则的支持度）：
  - $\mathrm{sup}(\mathrm{X} \rightarrow \mathrm{Y}) = \mathrm{P}(\mathrm{X} \cup \mathrm{Y}) = \frac{\mathrm{count}(\mathrm{X},\mathrm{Y})}{\mathrm{count}(\mathrm{Total})}$
    - $\mathrm{X} \cup \mathrm{Y}$ 看成一个事件，就是 X， Y 同时发生
- **Confidence**（置信度）：
  - $\mathrm{conf}(\mathrm{X} \rightarrow \mathrm{Y}) = \mathrm{P}(\mathrm{Y}|\mathrm{X}) = \frac{\mathrm{sup}(\mathrm{X} \cup \mathrm{Y})}{\mathrm{sup}(\mathrm{X})}$
    - 表示在包含 X 的 transaction 中，同时也包含 Y 的概率。
    - 就是算一个条件概率
- **Downward Closure Property**（向下封闭性质）：
  - 任意 frequent itemset 的所有子集也一定是 frequent 的。
- **Apriori Principle**：
  - 如果一个 itemset 是 infrequent 的，则它的所有超集也一定是 infrequent 的。

Association Rule 实例说明

以下是一个具体的例子，展示如何使用 Association Rule：

- **场景**：超市的市场篮子分析
  - 数据集包含顾客的购物记录，每条记录是顾客一次购物中购买的商品集合。
  - 目标：发现商品之间的关联关系，以优化商品摆放和促销策略。

- **步骤**：

    1. **生成频繁项集**：
        - 假设最小支持度阈值（minsup）为 50%。

```
Transaction ID | Items Bought
1              | {Milk, Bread, Butter}
2              | {Milk, Bread}
3              | {Milk, Butter}
4              | {Bread, Butter}
5              | {Milk, Bread, Butter}
```

规则 $\text{Milk} \rightarrow \text{Bread}$

1. **支持度计算**：
    - $\text{sup}(\text{Milk} \rightarrow \text{Bread}) = \frac{\text{包含 Milk,Bread 的 transaction 数量}}{\text{总 transaction 数量}}$
    - 包含 $\text{Milk}, \text{Bread}$ 的 transactions 是 {1, 2, 5}，数量为 3。
    - 总 transaction 数量为 5。
    - $\text{sup}(\text{Milk} \rightarrow \text{Bread}) = \frac{3}{5} = 60$
2. **置信度计算**：
    - $\text{conf}(\text{Milk} \rightarrow \text{Bread}) = \frac{\text{sup}(\text{Milk} \cup \text{Bread})}{\text{sup}(\text{Milk})}$
    - $\text{sup}(\text{Milk}) = \frac{\text{包含 Milk 的 transaction 数量}}{\text{总 transaction 数量}} = \frac{4}{5} = 80$
    - $\text{conf}(\text{Milk} \rightarrow \text{Bread}) = \frac{60}{80} = 75$

规则 $\text{Milk}, \text{Bread} \rightarrow \text{Butter}$

1. **支持度计算**：
    - $\text{sup}(\text{Milk}, \text{Bread} \rightarrow \text{Butter}) = \frac{\text{包含 Milk,Bread,Butter 的 transaction 数量}}{\text{总 transaction 数量}}$
    - 包含 $\text{Milk}, \text{Bread}, \text{Butter}$ 的 transactions 是 {1, 5}，数量为 2。
    - 总 transaction 数量为 5。
    - $\text{sup}(\text{Milk}, \text{Bread} \rightarrow \text{Butter}) = \frac{2}{5} = 40$
2. **置信度计算**：
    - $\text{conf}(\text{Milk}, \text{Bread} \rightarrow \text{Butter}) = \frac{\text{sup}(\text{Milk} \cup \text{Bread} \cup \text{Butter})}{\text{sup}(\text{Milk} \cup \text{Bread})}$
    - $\text{sup}(\text{Milk} \cup \text{Bread}) = \frac{\text{包含 Milk,Bread 的 transaction 数量}}{\text{总 transaction 数量}} = \frac{3}{5} = 60$
    - $\text{conf}(\text{Milk}, \text{Bread} \rightarrow \text{Butter}) = \frac{40}{60} = 66.7$

- 规则 $\text{Milk} \rightarrow \text{Bread}$ 的支持度为 60%，置信度为 75%。
- 规则 $\text{Milk}, \text{Bread} \rightarrow \text{Butter}$ 的支持度为 40%，置信度为 66.7%。

Apriori 算法总结

核心思想：

- 基于 **Downward Closure Property**（向下封闭性质）
    1. 任意 frequent itemset 的所有子集也一定是 frequent 的。
        - {A, B, C} frequent => {A, B}, {A, C}, {B, C} frequent
    2. 任意 infrequent itemset 的 超集（superset）也一定是 infrequent。可以不用生成。
        - {A, C} infrequent => {A, B, C} infrequent => 剪枝（停止生成）
- 通过逐层生成候选项集（candidate itemsets）并筛选频繁项集（frequent itemsets），减少搜索空间。
- **算法步骤**：
    1. **生成频繁 1-项集**：
        - 扫描数据库，计算每个单项的支持度（support），筛选出频繁 1-项集。
    2. **生成候选 k-项集**：
        - 通过频繁 (k-1)-项集的自连接（self-join）生成候选 k-项集。
        - **剪枝（pruning）**：移除包含 infrequent 子集的候选项集。
    3. **筛选频繁 k-项集**：
        - 扫描数据库，计算候选项集的支持度，筛选出频繁 k-项集。
    4. **重复步骤 2 和 3**，直到无法生成新的频繁项集。
    5. **生成关联规则**：
        - 对每个频繁项集生成所有可能的规则（rules），并计算置信度（confidence）。
        - 筛选出满足最小置信度阈值的规则。
- **优点**：
    - 简单易懂，基于直观的频繁项集生成过程。
    - 利用剪枝策略显著减少候选项集数量。
- **缺点**：
    - 需要多次扫描数据库，计算开销较大。
    - 候选项集数量可能呈指数增长，尤其在数据稠密或最小支持度较低时。
- **改进方法**：
    - **FP-Growth**：通过构建 FP 树避免候选项集生成。
    - **Eclat**：使用垂直数据格式加速计算。

Apriori 算法实例

以下是一个具体的例子，展示 Apriori 算法的步骤和原理：

数据集

假设有以下交易数据，最小支持度阈值（minsup）为 50%：

```
Transaction ID | Items Bought
1              | {Milk, Bread, Butter}
2              | {Milk, Bread}
3              | {Milk, Butter}
4              | {Bread, Butter}
5              | {Milk, Bread, Butter}
```

步骤

1. **生成频繁 1-项集**：
   - 计算每个单项的支持度：

     ```
     Item    | Support Count | Support
     Milk    | 4             | 80%
     Bread   | 4             | 80%
     Butter  | 4             | 8
     0%
     ```

   - 筛选出支持度 ≥ minsup 的项集：{Milk}, {Bread}, {Butter}。
2. **生成候选 2-项集**：
   - 通过频繁 1-项集的自连接生成候选 2-项集：

     ```
     {Milk, Bread}, {Milk, Butter}, {Bread, Butter}
     ```

   - 计算支持度：

     ```
     Itemset         | Support Count | Support
     {Milk, Bread}   | 3             | 60%
     {Milk, Butter}  | 3             | 60%
     {Bread, Butter} | 3             | 60%
     ```

   - 筛选出支持度 ≥ minsup 的项集：{Milk, Bread}, {Milk, Butter}, {Bread, Butter}。
3. **生成候选 3-项集**：
   - 通过频繁 2-项集的自连接生成候选 3-项集：

     ```
     {Milk, Bread, Butter}
     ```

   - 计算支持度：

```
Itemset               | Support Count | Support
{Milk, Bread, Butter} | 2             | 40%
```

- 筛选出支持度 ≥ minsup 的项集：无。

4. **生成关联规则**：
   - 对频繁项集生成规则并计算置信度：
     - 规则 $\text{Milk} \rightarrow \text{Bread}$：

```
Confidence = Support({Milk, Bread}) / Support({Milk}) = 60%
/ 80% = 75%
```

     - 规则 $\text{Milk, Bread} \rightarrow \text{Butter}$：

```
Confidence = Support({Milk, Bread, Butter}) / Support({Milk,
Bread}) = 40% / 60% = 66.7%
```

总结

- 频繁项集：{Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}。
- 关联规则示例：
  - $\text{Milk} \rightarrow \text{Bread}$，置信度为 75%。
  - $\text{Milk, Bread} \rightarrow \text{Butter}$，置信度为 66.7%。
