Covariance 定义、原理、计算与应用

- Covariance（协方差）是衡量两个变量之间共同变化趋势的统计量。
- 公式：$\mathrm{Cov}(X, Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$

原理：

- 协方差描述了两个变量如何共同变化：
- $\mathrm{Cov}(X, Y) > 0$：当 X 增加时，Y 也倾向于增加（正相关）。
- $\mathrm{Cov}(X, Y) < 0$：当 X 增加时，Y 倾向于减少（负相关）。
- $\mathrm{Cov}(X, Y) = 0$：X 和 Y 之间没有线性关系。

计算：

```
X = [1, 2, 3, 4, 5]
Y = [2, 4, 6, 8, 10]
```

- 计算均值：$\bar{X} = 3, \quad \bar{Y} = 6$
- 计算协方差：$\mathrm{Cov}(X, Y) = \frac{(1-3)(2-6)+(2-3)(4-6)+(3-3)(6-6)+(4-3)(8-6)+(5-3)(10-6)}{5}$

$$\mathrm{Cov}(X, Y) = \frac{20}{5} = 4$$

在数据分析中的应用：

1. **变量关系分析**：
   - 协方差用于判断两个变量之间是否存在正相关或负相关关系。
2. **特征选择**：
   - 在高维数据中，通过协方差矩阵分析变量之间的相关性，选择重要特征。
3. **主成分分析（PCA）**：
   - 协方差矩阵是 PCA 的基础，用于计算主成分方向。
4. **金融分析**：
   - 在投资组合中，协方差用于衡量资产之间的共同波动性，从而优化风险和收益。

Correlation 定义、原理与计算

- Correlation（相关性）是衡量两个变量之间线性关系强度和方向的统计指标。
  - 相关系数的取值范围为 $[-1, 1]$：
  - 1 表示完全正相关。
  - -1 表示完全负相关。
  - 0 表示无线性相关。
- 相关系数公式：$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$
  - 分子：协方差，表示两个变量的联合变化。
  - 分母：标准差的乘积，表示变量的独立变化。

示例：

```
X = [1, 2, 3, 4, 5]
Y = [2, 4, 6, 8, 10]
```

- 计算均值：$\bar{X} = 3, \quad \bar{Y} = 6$
- 计算相关系数：
  - $r = \frac{(1-3)(2-6)+(2-3)(4-6)+(3-3)(6-6)+(4-3)(8-6)+(5-3)(10-6)}{\sqrt{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2} \cdot \sqrt{(2-6)^2+(4-6)^2+(6-6)^2+(8-6)^2+(10-6)^2}}$
  - $r = \frac{20}{\sqrt{10} \cdot \sqrt{40}} = 1$
  - 结果：$r = 1$，表示完全正相关。

Lift 公式定义与使用

- Lift 衡量两个事件（itemsets）之间的相关性，表示它们是否独立或存在正/负相关关系。
- Lift 是条件概率与边缘概率的比值，用于衡量事件 X 和 Y 的关联强度：
- $\text{Lift}(X \to Y) = \frac{P(X \cup Y)}{P(X) \cdot P(Y)} = \frac{\text{conf}(X \to Y)}{P(Y)}$
  - $P(Y|X)$ 表示在事件 X 发生的条件下，事件 Y 发生的概率。
  - $P(Y)$ 表示事件 Y 的边缘概率。
- 解释：
  - 如果 $\text{Lift}(X \to Y) = 1$，则 X 和 Y 是独立的。
  - 如果 $\text{Lift}(X \to Y) > 1$，则 X 和 Y 是正相关的。
  - 如果 $\text{Lift}(X \to Y) < 1$，则 X 和 Y 是负相关的。

实例：

```
|            | Buy B | Not Buy B | Total |
|------------|-------|-----------|-------|
| Buy A      | 40    | 10        | 50    |
| Not Buy A  | 60    | 90        | 150   |
| Total      | 100   | 100       | 200   |
```

1. $P(A) = \frac{50}{200} = 0.25$
2. $P(B) = \frac{100}{200} = 0.5$
3. $P(A \cup B) = \frac{40}{200} = 0.2$
4. $\text{Lift}(A \to B) = \frac{P(A \cup B)}{P(A) \cdot P(B)} = \frac{0.2}{0.25 \cdot 0.5} = 1.6$

- 结果：$\text{Lift}(A \to B) = 1.6$，表示 A 和 B 之间存在正相关关系。

Chi-Square 定义与使用

- Chi-Square Test（卡方检验）是一种统计方法，用于检验两个变量之间是否存在显著的相关性。
- 通过比较观测值（Observed Value, O）与期望值（Expected Value, E）的差异，判断变量是否独立。
- **公式**：$\chi^2 = \sum \frac{(O-E)^2}{E}$
    - O：观测值。
    - E：期望值，通常根据独立假设计算：

期望值 (Expected Value) 的计算：

1. **独立性假设**：
    - 假设两个变量 A 和 B 是独立的，则联合概率 $P(A \cup B)$ 可以表示为边缘概率的乘积：
- $P(A \cup B) = P(A) \cdot P(B)$

2. 展开
    - $\frac{\text{count}(A \cup B)}{\text{Total}} = \frac{\text{count}(A)}{\text{Total}} \cdot \frac{\text{count}(B)}{\text{Total}}$
    - $E = \text{count}(A \cup B) = \frac{\text{count}(A) \cdot \text{count}(B)}{\text{Total}}$

- 期望值的计算基于独立性假设，反映了在变量独立情况下的理论频数。
- 通过比较观测值与期望值的差异，可以判断变量之间是否存在显著的相关性。

解释：

- 如果 $\chi^2 = 0$，表示观测值与期望值完全一致，变量独立。
- 如果 $\chi^2 > 0$，表示观测值与期望值存在差异，变量可能相关。
- 需要结合自由度（Degrees of Freedom, df）和显著性水平（Significance Level, α）查表判断是否显著。

示例：

```
```
	Buy B	Not Buy B	Total
Buy A	40	10	50
Not Buy A	60	90	150
Total	100	100	200
```
```

- 计算期望值：$E(\text{Buy A, Buy B}) = \frac{50 \times 100}{200} = 25$
- 计算 χ^2：$\chi^2 = \frac{(40-25)^2}{25} + \frac{(10-25)^2}{25} + \frac{(60-75)^2}{75} + \frac{(90-75)^2}{75} = 20$
- 根据自由度 $df = (2-1)(2-1) = 1$ 和显著性水平 $\alpha = 0.05$ 查表，判断是否显著相关。

Covariance, Correlation, Chi-Square, Lift 比较总结

Measure	Objective	Data Type	Output Range	Interpretation	Usage	Differences
Covariance	衡量两个变量的共同变化趋势	Numeric	$(-\infty, +\infty)$	正值：正相关；负值：负相关；0：无线性关系	用于变量关系分析、特征选择、主成分分析（PCA）、金融分析等	仅反映方向，不标准化，无法比较不同数据集的相关性
Correlation	衡量两个变量之间的线性关系强度和方向	Numeric	$[-1, 1]$	1：完全正相关；-1：完全负相关；0：无线性关系	用于变量关系分析、特征选择、模型评估等	标准化的协方差，适合比较不同数据集的相关性
Chi-Square	检验两个变量之间是否存在显著的相关性	Nominal (Categorical)	$[0, +\infty)$	0：完全独立；正值越大，相关性越强	用于分类变量的相关性检验、独立性检验等	仅适用于分类数据，需结合自由度和显著性水平判断
Lift	衡量两个事件（itemsets）之间的关联强度	Nominal (Categorical)	$(0, +\infty)$	1：独立；>1：正相关；<1：负相关	用于关联规则挖掘（如市场篮分析）	依赖于支持度和置信度，适合频繁模式挖掘

- **总结**：
  - **Covariance** 和 **Correlation** 适用于数值型数据，前者仅反映方向，后者标准化后便于比较。
  - **Chi-Square** 和 **Lift** 适用于分类数据，前者用于显著性检验，后者用于关联规则挖掘。
  - **Output Range** 和 **Interpretation** 是选择合适测度的关键。