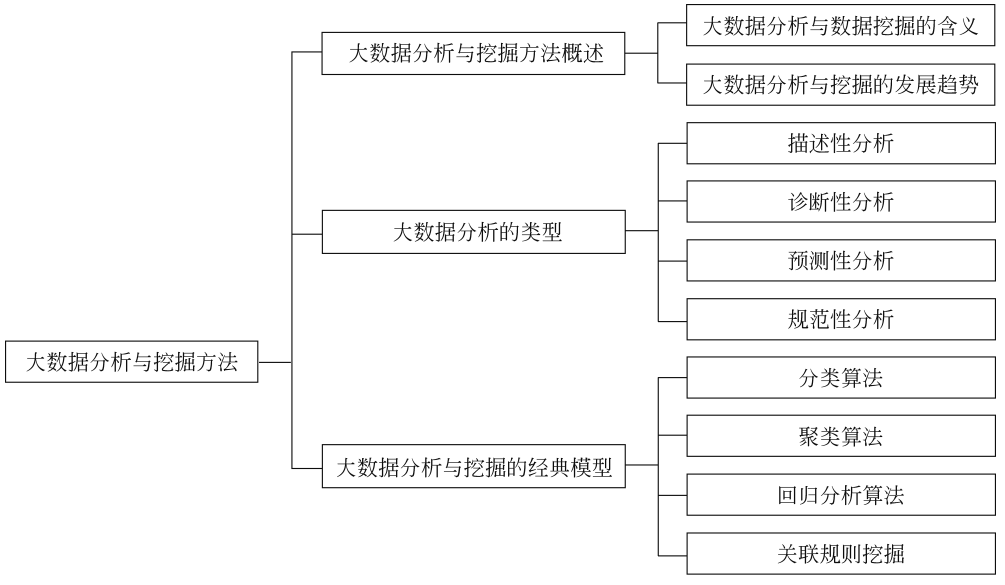
第 5 章

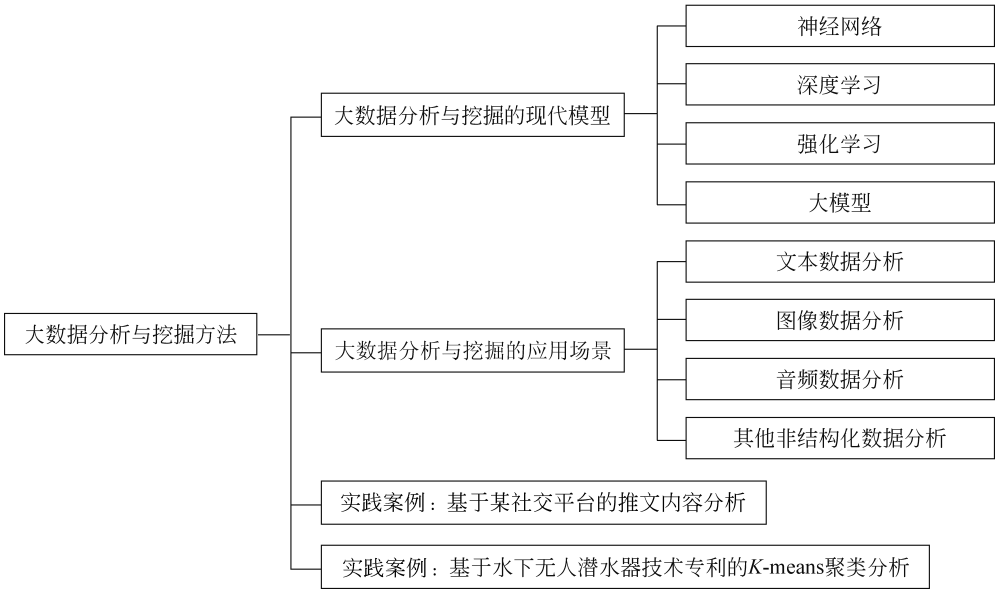
大数据分析与挖掘方法

在当今数字化时代，大数据分析与数据挖掘技术正以前所未有的速度发展，并深刻地改变着我们的生活和工作方式。随着互联网、物联网、移动互联网等技术的不断进步，数据的生成速度和数量呈指数级增长，数据已成为企业和组织中最宝贵的资产之一。大数据分析和挖掘技术应运而生，其核心目标是通过对大量数据进行处理、分析和挖掘，以发现隐藏在数据中的有价值信息和知识，从而支持决策和预测。目前，大数据分析与数据挖掘技术已经在生活中的各行各业得到了广泛应用。这些领域包括但不限于医疗健康、金融服务、零售业、交通运输和智能制造等。

本章要点：

- 大数据分析类型。
- 大数据分析与挖掘的经典模型。
- 大数据分析与挖掘的现代模型。
- 大数据分析与挖掘的应用场景。
- 实践案例：基于某社交平台的推文内容分析。
- 实践案例：基于水下无人潜水器技术专利的 K-means 聚类分析。





5.1　大数据分析与挖掘方法概述

5.1.1　大数据分析与数据挖掘的含义

大数据分析是指对规模巨大的数据进行分析，通过利用各种先进的分析技术和工具，从数据中提取有价值的信息和知识，以支持决策和预测。它涉及数据的收集、清洗、预处理、可视化、统计分析等多个步骤，以揭示数据中的模式、趋势、关联性及潜在价值。

数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含的、事先未知但潜在有用的信息和知识的过程。它融合了数据库技术、统计学、机器学习、人工智能等多个领域的知识和技术，运用统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法，旨在通过分析和建模发现数据中的模式、规律或关联性，从而支持决策制定和预测未来趋势。

5.1.2　大数据分析与挖掘的发展趋势

大数据分析与挖掘的发展趋势反映了技术进步、市场需求和行业变化的综合影响。以下是一些主要的发展趋势。

（1）**人工智能与机器学习的融合**：数据挖掘工具越来越多地集成人工智能和机器学习算法，以实现更先进的预测分析功能。这些技术使组织能够自动识别模式、生成可行的见解，并优化决策流程，从而提高效率和竞争力。

（2）**实时数据处理**：随着物联网和边缘计算的普及，实时数据挖掘变得越来越重要。数据挖掘工具正在整合实时分析功能，使组织能够在数据生成时进行分析并立即采取行动。这在金融、电子商务和物联网等行业尤其有价值。

（3）**自动化特征工程**：特征工程是数据挖掘过程中的关键步骤，涉及选择和创建与分析最相关的特征。自动化特征工程工具利用人工智能和机器学习算法从原始数据中自动识

别和创建相关特征，简化了数据准备过程，减少了人工工作量，并提高了数据挖掘模型的准确性。

（4）**基于云的数据挖掘解决方案**：基于云的数据挖掘解决方案的采用率正在上升，这得益于其提供的可扩展性、灵活性和成本效益。云解决方案使组织能够访问强大的分析工具和资源，而无需本地基础设施。此外，云解决方案促进跨团队和跨地点的协作和数据共享，使组织更轻松地利用数据挖掘的优势。

（5）**增强的数据可视化**：数据可视化在数据挖掘中发挥着至关重要的作用，使用户能够理解复杂的数据集并有效地传达见解。数据挖掘工具整合了先进的数据可视化功能，如交互式仪表板、热图和地理空间可视化，帮助用户更直观地探索和解释数据。

（6）**模型可解释性**：模型可解释性工具变得越来越重要，使用户能够了解数据挖掘模型如何进行预测并信任它们产生的结果。这有助于提高决策的透明度和可信度。

（7）**大数据技术的集成**：Hadoop、Spark 和 NoSQL 数据库等大数据技术的激增正在推动数据挖掘工具与这些平台的集成。数据挖掘工具利用大数据技术的可扩展性和处理能力，有效地分析大型且多样化的数据集，支持跨行业的数据驱动决策。

大数据分析与挖掘工具市场正在不断发展，以满足对高级分析、实时洞察和道德数据使用不断增长的需求。通过拥抱人工智能与机器学习的融合、实时数据处理、自动化特征工程、基于云的数据挖掘解决方案、增强的数据可视化、模型可解释性与大数据技术的集成以及对隐私和道德数据挖掘的关注，组织可以利用这些工具释放数据的全部潜力，推动数字时代的创新和增长。

5.2 大数据分析的类型

5.2.1 描述性分析

在实际工作中，对于刚接手的数据集，在正式处理需求前，除了需要梳理清楚数据上报及转发环节，还需要对数据集进行质量评估和了解数据集的描述性统计特征。分析师尤其需要重点关注数据集的“描述性统计特征”，了解不同类型的数据的集中、离散和分布程度，以便在业务提数时，补充合理的筛选条件，避免计算出来的指标有误导性。

“描述性统计”是以数字和图表的形式来理解、分析和总结数据，它是数据分析的第一步，通常在数据收集完成后进行，有助于研究者更好地理解数据、发现异常值、探索数据的特征与趋势，为下一步的统计推断与建模做准备。

常见的描述性统计量包括数据的集中趋势描述、数据的离中趋势描述、数据的分布形态描述、数据的分布特征可视化。

1. 数据的集中趋势描述

集中趋势（Central Tendency）反映的是一组数据向某一中心值靠拢的倾向，在中心值附近的数据数目较多，而远离中心值的数据数目较少。对集中趋势进行描述就是寻找数据一般水平的中心值或代表值。了解数据集中的趋势对揭示数据的总体行为至关重要。常用度量指标有众数、中位数、平均数等。

（1）众数。众数（Mode）是指在统计分布上具有明显集中趋势点的数值，代表数据的一

般水平，也是一组数据中出现次数最多的数值，有时众数在一组数中有好几个。在统计实践中，常利用众数来近似反映社会经济现象的一般水平。

（2）中位数。中位数也称为中值（Median），是指将统计总体当中的各个变量值按大小顺序排列起来，形成一个数列，处于数列中间位置的变量值就称为中位数。中位数能够避免数据的平均水平受到异常值的影响，因此在做数据分析时，不仅要计算算术平均数，也要计算中位数，若两个数字差距很大，就用中位数作为平均数。

（3）平均数。平均数包括算术平均数、几何平均数和调和平均数。

- 算术平均数。算术平均数（Arithmetic Mean）是总体中各个体的某个数量标志的总和与个体总数的比值。算术平均数是集中趋势中最主要的测度值，但是容易受极端数值的影响，当数据集合中有极大值或极小值存在时，会对算术平均数产生很大的影响，计算结果会掩盖数据集合的真实特征。它的基本公式是

$$A_n=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}$$

- 几何平均数。几何平均数（Geometric Mean）是 n 个变量值连乘积的 n 次方根。几何平均数是计算平均比率和平均速度最适用的一种方法。有些数据之间的关系不是相加减关系，而是乘除关系，此时应该用几何平均数来表示由这样的数值组成数据集合的集中趋势。通常用于计算“率”和“指数”，比如增长率、增长指数等。

$$G_n=\sqrt[n]{x_1x_2\cdots x_n}=\sqrt[n]{\prod_{i=1}^{n}x_i}$$

- 调和平均数。在统计分析中，有时会由于种种原因没有频数的资料，只有每组的变量值和相应的标志总量。这种情况下就不能直接运用算术平均方法来计算了，而需要以迂回的形式，即用每组的标志总量除以该组的变量值推算出各组的单位数，才能计算出平均数，我们可以用调和平均的方法完成这个计算。

调和平均数（Harmonic Mean）是各变量值倒数的算术平均数的倒数。由于它是根据变量值倒数计算的，所以又称作倒数平均数。如果一组数据中较大的数值把算术平均数拉高了，但我们想放大较小值的影响，减小较大值的影响，就可以使用调和平均数。

$$H_n=\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

2. 数据的离中趋势描述

离中趋势是数据分布的一个重要特征，它反映各变量值远离其中心值的程度。离中趋势从侧面说明了集中趋势测度值的代表程度，数据的离中趋势越大，集中趋势的测度值对该组数据的代表性就越差；数据的离中趋势越小，集中趋势的测度值的代表性就越好。描述离中趋势的统计指标主要有方差、标准差、极差、四分位数。

（1）方差。方差（Variance）是一组数据的平均值与每个数据点的差值的平方和的平均值，它可以反映一组数据的离散程度。如果一组数据的方差较小，则表明这组数据的分布较为集中；反之，如果方差较大，则表明这组数据的分布较为分散。对未经分组的数据资料，总体方差计算公式如下：

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

方差有总体方差与样本方差之分，上面说的是总体方差，如果要计算样本方差，只需要在分母上减一。一般把样本方差记为 S^2。

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

(2) 标准差。标准差(Standard Deviation)又称均方差，它是各单位变量值与其平均数离差平方的平均数的方根，通常用 σ 表示。它是测度数据离散程度的最主要方法。标准差是具有量纲的，它与变量值的计量单位相同，标准差的本质是求各变量值与其平均数的距离和，即先求出各变量值与其平均数离差的平方，再求其平均数，最后对其开方。之所以称其为标准差，是因为在正态分布条件下，它和平均数有明确的数量关系，是真正度量离中趋势的标准。

总体标准差 σ 的计算公式如下：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

标准差同样有总体标准差与样本标准差之分，样本标准差 S 的计算公式如下：

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

方差/标准差值的增大表示数据点相对其平均值的差异增大，即数据的离散程度增大。在风险评估模型的处理中，数据的波动性越高，意味着它所包含的信息量越大。信息量的增加意味着模型需要处理更多不确定因素，因此相应的风险也会增加。

(3) 极差。极差也称为全距(Range)，指一组数据中最大值和最小值之间的差值，计算简单，易于理解，体现了数据的波动范围。但因其很容易受到极端值的影响，不能反映中间数据的分散情况，需要结合其他离散程度描述指标来描述数据集合的离散程度。

$$R = x_{\max} - x_{\min}$$

(4) 四分位差。四分位数(Quartile)是指在统计学中把所有数值由小到大排列并分成四等份(每等份包含 25%的数据)，处于三个分割点位置的数值。四分位数有 3 个，第 1 个四分位数称为下四分位数，即 25%分位数；第 2 个四分位数就是中位数，即 50%分位数；第 3 个四分位数称为上四分位数，即 75%分位数，分别用 Q_1、Q_2、Q_3 表示。

四分位差(Quartile Deviation)又称四分位距，其值等于第 1 个四分位数与第 3 个四分位数的差值($Q_3 - Q_1$)。四分位差反映了中间 50%数据的离散程度。其数值越小，说明中间的数据越集中；数值越大，说明中间的数据越分散。四分位差不受极端值影响，因此，在某种程度上弥补了极差的一个缺陷。

$$Q_d = Q_3 - Q_1$$

四分位数与四分位差示意如图 5.1 所示。

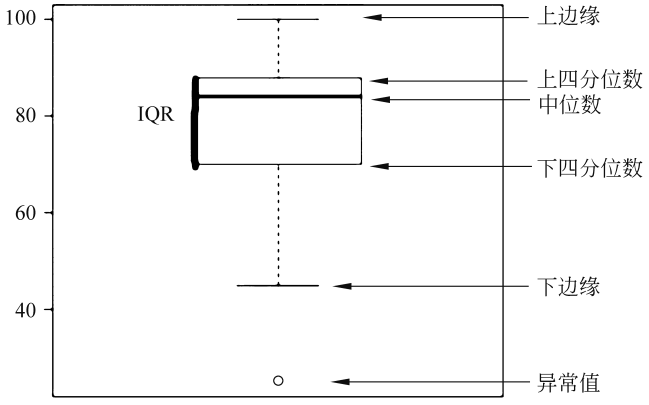


图 5.1　四分位数与四分位差示意

3. 数据的分布形态描述

集中趋势和离散程度是数据分布的两个重要特征，尤其是均值和标准差。对于正态分布，只要知道了均值和标准差，就可以确定其分布。但对于未知的分布，要想全面了解数据分布的特点，不仅要掌握数据的集中趋势和离散程度，还需要知道数据分布的形状是否对称、偏斜的程度以及分布的扁平程度等，这些统称为分布的形态。

数据的分布形态是指以数据大小为 x 轴，数据数量为 y 轴，展示数据的排列方式和模式，偏态和峰态就是对分布形态的测度。其中，“偏态”(Skewness)是对数据分布对称性的测度，“峰态”(Kurtosis)是对数据分布平峰或尖峰程度的测度，如图 5.2 所示。

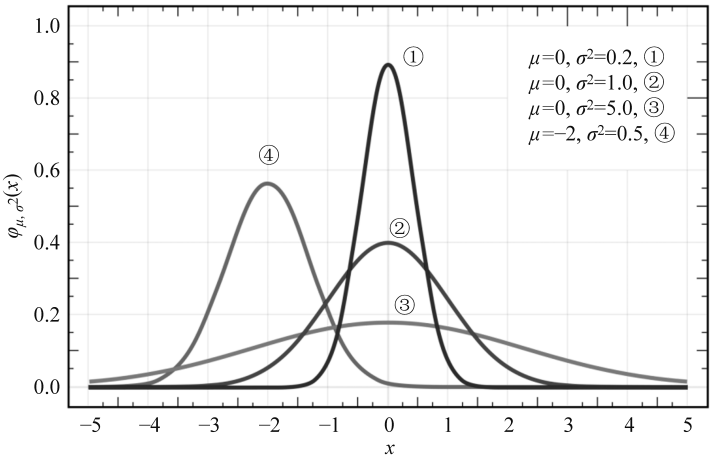


图 5.2　数据的分布形态

(1) 偏态。偏态是对分布偏斜方向和程度的测度，有些变量值出现的次数往往是非对称型的，如收入分配、市场占有份额、资源配置等。变量分组后，总体中各个体在不同的分组变量值下分布并不均匀对称，而呈现出偏斜的分布状况，统计上将其称为偏态分布。

利用众数、中位数和平均数之间的关系就可以判断分布是对称、左偏还是右偏，但要测度偏斜的程度则需要计算偏态系数(Coefficient of Skewness)。统计分析中测定偏态系数的方法很多，一般采用矩的概念计算，其计算公式为三阶中心矩 μ_3 与标准差的三次方之比。具体公式如下：

$$S_k=\frac{\mu_3}{\sigma^3}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{\frac{3}{2}}}$$

偏态系数可以帮助我们了解数据集是否向某一方向倾斜。一般情形下，当分布对称时，$S_k=0$；分布不对称时，$S_k<0$ 为左偏（或负偏）分布，$S_k>0$ 为右偏（或正偏）分布。偏态系数绝对值越大，分布形态的偏斜程度越大。左偏、对称与右偏分布示意如图 5.3 所示。

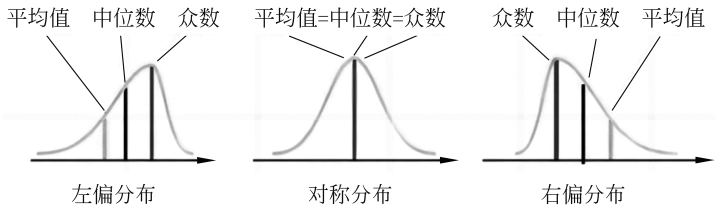


图 5.3　左偏、对称与右偏分布示意

在对称分布的情况下，一般是众数＝中位数＝平均值，但是在现实生活中，有些获取到的数据并不是对称分布的，会产生数据偏态的情况；就会用偏态来描述数据分布的形状是否对称，及偏斜的程度。

（2）峰态。峰态是分布集中趋势高峰的形状。在变量数列的分布特征中，常常以正态分布为标准，变量数列分布曲线顶峰的尖平程度在统计上称为峰态。如果分布的形状比正态分布更高更尖，则称为尖峰分布；如果分布的形状比正态分布更矮、更胖，则称为平峰分布，如图 5.4 所示。

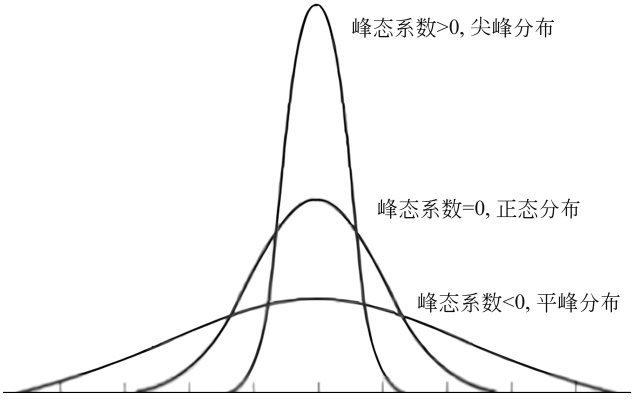


图 5.4　峰态分布示意图

峰态系数（Coefficient of Kurtosis）是统计中描述次数分布状态的又一个重要特征值，用以测定邻近数值周围变量值分布的集中或分散程度。一般采用矩的概念计算，即运用四阶中心矩 μ_4，与标准差的四次方对比，以此来判断各分布曲线峰态的尖平程度。峰态系数计算公式如下：

$$K=\frac{\mu_4}{\sigma^4}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\sigma^4}-3=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}-3$$

由于正态分布的峰态系数为0，当$K<0$时，表示该总体数据分布与正态分布相比较为扁平，数据的分布更分散；当$K>0$时，表示该总体数据分布与正态分布相比较为陡峭，数据的分布更集中。

4. 数据的分布特征可视化

除了数值度量，可视化也是一种强大的手段。直方图、箱线图和概率图等图形工具可以帮助我们更直观地理解数据的分布特征。通过可视化，我们能够捕捉到数据中的模式、异常值和趋势。此部分将在后续章节中详细展开介绍。

5.2.2　诊断性分析

诊断性分析的主要目的是通过深入挖掘数据的背后原因，解释数据异常或变动的原因。与描述性分析关注数据的总体概况不同，诊断性分析侧重于对特定问题、异常或趋势的深入理解，以识别导致这些现象的根本原因，并提供解决问题的策略。简单来说，诊断性分析的目标是探究原因，解释为什么某些事情会发生。它关注的是“为什么会发生这样的结果”。常见的诊断性分析包括比较分析、相关分析和卡方检验等。

1. 比较分析

比较分析又称对比分析，是一种将两个或两类事物的市场资料或数据相比较，以确定它们之间相同点和不同点的逻辑方法。这种方法常用于社会科学研究中，尤其在经济学、管理学、社会学等领域有着广泛的应用。通过对比不同对象，揭示其异同点，可以为决策者提供科学依据，帮助其作出正确评价。比较分析不仅限于简单的数据对比，它更注重通过对比揭示事物的本质和规律。这种方法通过对比不同对象在特定方面的表现，帮助人们更深入地理解事物之间的差异和联系，从而为决策提供有力支持。

2. 相关分析

相关分析是指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度。常见的用于相关分析的相关系数有三种：Pearson、Spearman和Kendall。现实场景中使用Pearson相关系数的情况比较多。三种相关系数的区别如表5.1所示。

表5.1　三种相关系数的区别

相关系数	适用场景	备　注
Pearson	定量数据，数据满足正态分布	正态图可查看正态性，散点图展示数据关系
Spearman	定量数据，数据不满足正态分布	通过*Q-Q*图或统计检验验证正态性
Kendall	定量数据一致性判断	通常用于评分数据一致性水平研究（非关系研究），如评委打分、数据排名等

相关系数r的取值范围为$[-1,1]$，负值表示负相关，正值表示正相关，0表示无相关性。r值大小与关系强度判定如表5.2所示。

表5.2　r值大小与关系强度判定

相关系数	$\|r\|>0.95$	$\|r\|\geqslant 0.8$	$0.5\leqslant\|r\|<0.8$	$0.3\leqslant\|r\|<0.5$	$\|r\|<0.3$
关系强度	显著相关	高度相关	中度相关	低度相关	弱相关

3. 卡方检验

卡方检验(Chi-Squared Test 或 χ^2 Test)是一种统计量的分布在零假设成立时近似服从卡方分布(χ^2 分布)的假设检验,用于确定两个分类变量之间是否存在关联性。它通过比较观察值和期望值之间的差异来评估变量之间的独立性。在没有其他的限定条件或说明时,卡方检验一般代指的是皮尔森卡方检定。卡方检验的基本公式(其中,A 为实际频数,T 为理论频数,χ 为卡方值如下):

$$\chi^2 = \sum \frac{(A-T)^2}{T}$$

5.2.3 预测性分析

预测性分析(Predictive Analysis)是一种数据分析方法,通过构建和应用预测模型,利用历史数据和当前数据来预测未来的事件或趋势。这些模型可以基于统计学、机器学习、数据挖掘等技术,帮助企业和组织提前了解未来可能发生的情况,从而制定相应的策略和计划。

预测性分析的核心在于构建模型,这包括了建立和验证提供精准预测的模型,使现有的数据被理解为推断未来的事件或简单地预测未来的数据。预测性分析是一个复杂的领域,它需要海量的数据支撑,以及对预测模型的熟练运用和精细调整,以确保分析结果的准确性。常见的预测性分析包括回归分析、时间序列分析等。预测性分析常常依赖于机器学习的算法,如随机森林、支持向量机等,以及用于学习和测试数据的统计数据。

1. 回归分析

在统计学中,回归分析(Regression Analysis)指的是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。回归分析按照涉及的变量的多少,可分为一元回归分析和多元回归分析;按照因变量的多少,可分为简单回归分析和多重回归分析;按照自变量和因变量之间的关系类型,可分为线性回归分析和非线性回归分析。在大数据分析中,回归分析是一种预测性的建模技术,它研究的是因变量(目标)和自变量(预测器)之间的关系。这种技术通常用于预测性分析、时间序列模型以及发现变量之间的因果关系。

(1) 线性回归。它是最为人熟知的建模技术之一。线性回归通常是人们在学习预测模型时首选的技术之一。在这种技术中,因变量是连续的,自变量可以是连续的,也可以是离散的,回归线的性质是线性的。线性回归使用最佳的拟合直线(也就是回归线)在因变量(Y)和一个或多个自变量(X)之间建立一种关系。

(2) 逻辑回归。逻辑回归根据给定的自变量数据集来估计事件的发生概率,由于结果是一个概率,因此因变量的范围为[0,1]。逻辑回归的因变量可以是二分类的,也可以是多分类的,但是二分类的更为常用,也更加容易解释,因此,实际中最为常用的就是二分类的逻辑回归。

2. 时间序列分析

时间序列分析(Time-Series Analysis)是指将原来的数据分解为四部分——趋势、周期、时期和不稳定因素,然后综合这些因素,提出预测。

时间序列分析是定量预测方法之一。它包括一般统计分析(如自相关分析、谱分析等)、统计模型的建立与推断,以及关于时间序列的最优预测与控制等内容。经典的统计分析都

假定数据序列具有独立性，而时间序列分析则侧重研究数据序列的互相依赖关系。后者实际上是对离散指标的随机过程的统计分析，所以又可看作随机过程统计的一个组成部分。例如，记录了某地区第1个月，第2个月……第 N 个月的降雨量，利用时间序列分析方法，可以对未来各月的雨量进行预报。

时间序列分析主要有确定性变化分析和随机性变化分析。其中，确定性变化分析包括趋势变化分析、周期变化分析、循环变化分析。随机性变化分析则包括 AR、MA、ARMA、ARIMA 模型等。

5.2.4　规范性分析

规范性分析(Prescriptive Analysis)是一种数据分析方法，它是预测性分析的下一步，旨在通过大量使用人工智能分析数据，来识别可用于进行预测和确定最佳行动方案的模式。它不仅预测未来事件，还提出了实现预期结果的最佳行动方案。企业可利用规范性分析来完成多项任务，如客户细分、流失率预测、欺诈检测、风险评估、需求预测、规范性维护、个性化推荐等。

虽然规范性分析实践在大数据诞生之前便已存在，但是随着各行各业的企业积累大量历史数据，这种分析实践得到了加速应用。如今，规范性分析工具使用预测建模中的许多统计技术，但也利用人工智能和机器学习算法及模型。分析软件使用经过大量数据训练的机器学习模型，使分析师能够更准确地识别风险和机会，从而指导和改进业务领导者的决策。规范性分析依赖于优化和基于规则的决策技术。其中的数学模型包括自然语言处理、机器学习、统计、运筹学等。常用的规范性分析有线性规划、决策树分析等。

1. 线性规划

线性规划(Linear Programming，LP)是运筹学中研究较早、发展较快、应用广泛、方法较成熟的一个重要分支，是辅助人们进行科学管理的一种数学方法，是研究线性约束条件下线性目标函数的极值问题的数学理论和方法。线性规划广泛应用于军事作战、经济分析、经营管理和工程技术等方面，可为合理地利用有限的人力、物力、财力等资源作出最优决策提供科学的依据。

线性规划问题通常可以表示为以下标准形式。

目标函数：

$$\text{Maximize}=c_1 \cdot x_1+c_2 \cdot x_2+\cdots+c_n \cdot x_n \quad \text{或} \quad \text{Minimize}=c_1 \cdot x_1+c_2 \cdot x_2+\cdots+c_n \cdot x_n$$

约束条件：

$$\begin{aligned} &a_{11} \cdot x_1+a_{12} \cdot x_2+\cdots+a_{1n} \cdot x_n \leqslant b_1 \\ &a_{21} \cdot x_1+a_{22} \cdot x_2+\cdots+a_{2n} \cdot x_n \leqslant b_2 \\ &\cdots \\ &a_{m1} \cdot x_1+a_{m2} \cdot x_2+\cdots+a_{mn} \cdot x_n \leqslant b_m \end{aligned}$$

非负约束：

$$x_1, x_2, \cdots, x_n \geqslant 0$$

其中，$c_1, c_2, \cdots, c_n$ 是目标函数的系数，a_{ij} 是约束条件的系数，$b_1, b_2, \cdots, b_m$ 是约束条件的右侧常数，$x_1, x_2, \cdots, x_n$ 是决策变量。