frlily / February 07, 2025 09:38AM
“抽樣”在資料分析中的應用

1. 1. 01 大資料還要抽樣？
2. 2. 02 常見的抽樣方法
3. 3. 03 過取樣 vs 欠取樣
4. 4. 04 如何正確使用抽樣

以下文章來源於數據分析星球 ，作者數據分析星球

1. 01 大資料還要抽樣？

在資料分析領域，資料抽樣是一個非常重要的概念。資料抽樣指的是從整個資料集合中選取一部分資料進行分析，這樣可以使資料分析更加高效和精確。本文將從以下幾個方面來介紹資料抽樣的相關內容。

1 什麼是資料抽樣？

資料抽樣指的是從整個資料集合中選取一部分資料進行分析。資料抽樣可以減少資料分析的成本和時間，同時也可以使資料分析結果更加精確和可靠。在資料抽樣的過程中，要注意選擇合適的抽樣方法和樣本量，以保證抽樣結果的代表性。

2 大資料時代還需要抽樣麼？

在大資料時代，資料量的增長迅速，資料分析也變得更加複雜。因此，抽樣在大資料分析中仍然是非常重要的。在大資料分析中，透過對資料進行抽樣，可以使分析結果更加精確，同時也可以減少資料分析的成本和時間。

2. 02 常見的抽樣方法

簡單隨機抽樣

簡單隨機抽樣是一種簡單的抽樣方法，它是從整個資料集合中隨機選取一定數量的樣本進行分析。這種方法適用於資料分佈均勻的情況下，每個樣本被選中的機率相等。

舉例來說，我們想要對一個市場上的商品進行價格調研，我們可以透過簡單隨機抽樣的方法，從所有商品中隨機選取一定數量的商品進行價格調研。

系統抽樣

系統抽樣是一種有規律的抽樣方法，它是從整個資料集合中按照一定的規律選取樣本進行分析。這種方法適用於資料分佈不均勻的情況下。

例如，我們想要對一家公司進行員工滿意度調查，我們可以透過系統抽樣的方法，按照公司的部門結構，每隔一定數量的員工進行抽樣，以保證樣本具有代表性。

分層抽樣

分層抽樣是一種按照資料分層的抽樣方法，它是將資料集合分為多個層次，然後在每個層次中按照一定的規則選取樣本進行分析。這種方法適用於資料分佈不均勻，並且資料可以按照某種規則劃分為多個層次的情況下。

例如，一家公司有3個部門，想要對每個部門的員工進行薪資調查，可以使用分層抽樣方法進行抽樣。

整群抽樣

整群抽樣適用於樣本資料呈現群體結構的情況下。例如，一條生產線上的產品按照批次分為多個群體，想要對每個群體進行抽樣檢驗，可以使用整群抽樣方法進行抽樣。

3. 03 過取樣 vs 欠取樣

在機器學習中，為了使模型更加準確，有時需要對資料進行抽樣處理。過取樣和欠取樣是抽樣過程中常見的問題。

過取樣指的是在樣本中出現了一些資料過多的類別，而另一些類別的資料卻較少的情況。這會導致模型過分關注某些類別，從而降低整體預測效果。解決過取樣的方法有兩種，一種是增加欠取樣類別的樣本，另一種是減少過取樣類別的樣本。

欠取樣指的是在樣本中某些類別資料較少，而另一些類別資料較多的情況。這會導致模型對資料的刻畫不夠全面，從而降低整體預測效果。解決欠取樣的方法有兩種，一種是減少過取樣類別的樣本，另一種是增加欠取樣類別的樣本。

4. 04 如何正確使用抽樣

在資料分析中，抽樣方法是非常常見的技術，正確使用抽樣方法可以提高資料分析的準確性和效率。以下是一些使用抽樣方法的建議：

4.1 確定目標

在使用抽樣方法之前，首先需要明確分析的目標，確定要分析的特徵和指標。這樣可以幫助確定取樣的樣本數量和取樣方法。

4.2 確定取樣方法

根據分析的目標和資料的特點，選擇適當的抽樣方法。例如，如果資料集比較大且分佈均勻，可以選擇簡單隨機抽樣；如果資料集包含多個層次，可以選擇分層抽樣等。

4.3 確定樣本數量

確定樣本數量需要考慮多方面因素，例如資料集的大小、樣本的分佈、取樣方法等。通常，樣本數量需要滿足一定的置信度和置信區間要求，以保證資料分析的可靠性和準確性。

4.4 驗證抽樣結果

在使用抽樣方法後，需要對結果進行驗證。可以使用隨機抽樣或重複抽樣的方法來驗證結果的可靠性和準確性。