

frilly / November 27, 2024 11:35AM

[資料分析中常用的“迴歸演算法”及其應用](#)

- [1. 1.01 什麼是迴歸？](#)
- [2. 2.02 線性迴歸演算法](#)
- [3. 3.03 多項式迴歸演算法](#)
- [4. 4.04 決策樹迴歸演算法](#)
- [5. 5.05 隨機森林迴歸](#)
- [6. 6.06 總結](#)

以下文章來源於數據分析星球，作者數據分析星球

1.01 什麼是迴歸？

隨著計算機技術的不斷髮展，機器學習在各行各業中得到了廣泛的應用。迴歸演算法是機器學習中的重要部分，它可以用於預測數值型的輸出結果。在本文中，我們將介紹迴歸演算法的幾種分類：線性迴歸、多項式迴歸、決策樹迴歸和隨機森林迴歸等，並透過生活中的案例用大白話講解什麼是迴歸。

在生活中，我們經常需要根據已知資料來預測未知資料的結果。例如，我們可以通過歷史房價和房屋面積的關係來預測一套新房子的價格。這個預測過程就是迴歸。在機器學習中，迴歸就是根據已知資料來預測數值型的輸出結果。

下面我們將介紹幾種常見的迴歸演算法，並用簡單的語言描述演算法原理，最後我們也簡單地用Python實現各個迴歸演算法。

2.02 線性迴歸演算法

線性迴歸是一種常見的迴歸演算法，它的原理是透過尋找最佳的直線來擬合資料。這條直線可以用來預測新的資料點。在線性迴歸中，我們可以透過最小二乘法來確定最佳的擬合直線。

演算法原理

線性迴歸是一種用於預測連續變數的統計學習演算法。其基本原理是利用輸入特徵與輸出變數之間的線性關係來建立一個預測模型。該模型可以表示為一個線性方程，其中輸入特徵是自變數，輸出變數是因變數。透過訓練資料集來確定這個線性方程中的係數，即可以得到一個能夠對新的輸入資料進行預測的模型。

線性迴歸的核心是求解最小二乘法。最小二乘法的目標是使模型預測值與真實值之間的誤差平方和最小化，即透過最小化殘差平方和來獲得最佳的擬合直線。為了實現這個目標，我們使用梯度下降演算法來最佳化損失函式，即最小化殘差平方和。

案例分析

使用Python中的scikit-learn庫中的線性迴歸演算法預測房價。

假設我們有一組房屋面積和價格的資料，我們可以使用線性迴歸演算法來預測房價。我們使用Python中的scikit-learn庫來實現。

首先，我們需要匯入必要的庫並讀取資料：

然後，我們可以使用scikit-learn庫中的LinearRegression類來訓練模型並進行預測：

3. 03 多項式迴歸演算法

多項式迴歸演算法是一種基於線性迴歸的演算法，在線性迴歸基礎上增加多項式特徵來擬合非線性關係。

演算法原理

如果一些資料點不能透過一條直線很好地擬合，也就是說並不是線性關係，我們可以透過新增二次項來擬合這些資料點。多項式迴歸可以透過新增任意次冪的項來實現，可以更好地擬合非線性關係。

多項式迴歸演算法的原理是將每個特徵的冪次方作為一個新的特徵，然後使用線性迴歸演算法來擬合這些特徵。在實踐中，我們可以使用Scikit-learn庫中的PolynomialFeatures類來建立多項式特徵。在這個類中，我們可以指定多項式的最高次數，然後將原始特徵轉換為多項式特徵，最後使用線性迴歸演算法來訓練模型。

案例分析

我們以一個銷售量預測為例，假設我們有以下資料：

資料中包含了每個產品的廣告費用和對應的銷售量。我們希望透過多項式迴歸演算法預測未來的銷售量。首先，我們使用PolynomialFeatures類將原始特徵轉換為二次多項式特徵：

在上面的程式碼中，我們使用了Pandas庫來載入資料，並將廣告費用和銷售量分別儲存在X和y中。然後，我們建立一個PolynomialFeatures物件，將X轉換為二次多項式特徵，並將結果儲存在X_poly中。接下來，我們使用LinearRegression類來訓練模型，並使用predict()方法預測新的銷售量。

4. 04 決策樹迴歸演算法

決策樹迴歸演算法是一種基於決策樹的迴歸演算法。與分類問題類似，決策樹迴歸也是建立一棵樹來預測輸出結果。不同之處在於，分類問題的決策樹每個葉子節點表示一個類別，而迴歸問題的決策樹每個葉子節點表示一個數值。

演算法原理

決策樹迴歸演算法的基本原理是透過選擇最佳的劃分屬性和劃分點將資料集劃分為多個子集，使得每個子集內資料的輸出結果儘可能相似。劃分屬性和劃分點的選擇是透過計算每個屬性對資料集的資訊增益或資訊增益比來完成的。

資訊增益是指在已知某個屬性的情況下，資料集的無序性減少的程度。資訊增益越大，說明使用該屬性劃分可以使得資料集的無序性減少得更多。資訊增益比則是在資訊增益的基礎上，考慮屬性本身帶來的資訊量，避免了選擇取值數目較多的屬性的傾向。

當決策樹建立完成後，對於新的樣本，透過從根節點開始依次比較該樣本的屬性值和決策樹中的劃分點，直到找到相應的葉子節點。葉子節點的輸出結果即為該樣本的預測值。

案例分析

本案例將使用決策樹迴歸演算法預測股票價格。資料集包含了20個股票的基本面資料和對應的股票價格。

首先，載入資料集並劃分為訓練集和測試集。

接著，使用scikit-learn庫中的DecisionTreeRegressor類建立決策樹迴歸模型，並訓練模型。

執行結果就可以計算出r2_score，透過該指標可以說明決策樹迴歸演算法對該資料集的預測效能的優劣。

5.05 隨機森林迴歸

隨機森林迴歸是一種基於決策樹的整合學習演算法，它是由多個決策樹組成的模型，透過對每個決策樹的預測結果進行平均或加權平均，得到最終的預測結果。隨機森林是一種強大的迴歸演算法，它具有高準確度、抗噪聲能力強、對缺失資料不敏感、可處理大量資料等優點，因此被廣泛應用於各個領域的預測任務。

演算法原理

隨機森林迴歸演算法的原理與決策樹迴歸演算法類似，不同之處在於隨機森林中使用的是多個決策樹，且每個決策樹都是在隨機樣本和隨機特徵的基礎上構建的。具體地，隨機森林的構建過程如下：

- 從樣本集中隨機抽取一定數量的樣本，作為訓練集。
- 從特徵集中隨機選擇一部分特徵，用於構建決策樹。
- 構建多個決策樹，每個決策樹都是在上述步驟的基礎上構建的。
- 對於新的樣本，使用每個決策樹進行預測，得到多個預測結果。
- 對多個預測結果進行平均或加權平均，得到最終的預測結果。

案例分析

假設我們有過去幾年的氣溫資料，我們希望根據這些資料來預測未來一週的氣溫。首先，我們需要匯入資料並對資料進行預處理，包括資料清洗、特徵選擇和特徵縮放等。這裡我們使用Python中的pandas和scikit-learn庫來實現資料處理過程。

這段程式碼首先匯入需要的庫和模組，然後使用Pandas庫讀取氣溫資料。接著，我們將資料分離為特徵和標籤，並使用train_test_split函式將資料劃分為訓練集和測試集。然後，我們使用RandomForestRegressor類構建隨機森林迴歸模型，並將模型訓練於訓練集上。最後，我們使用模型在測試集上進行預測，並使用mean_squared_error函式計算預測結果和真實結果之間的均方根誤差(RMSE)，透過這些指標表徵模型的預測效能。

6.06 總結

透過本文的介紹，我們瞭解了機器學習中的迴歸演算法，包括線性迴歸、多項式迴歸、決策樹迴歸和隨機森林迴歸。

這些演算法可以透過訓練模型來預測未來的數值型資料，因此在許多實際應用中被廣泛使用。

在實際使用這些演算法時，需要根據資料特點和應用場景選擇不同的演算法。線性迴歸適用於簡單的線性關係，而多項式迴歸可以處理更複雜的曲線關係。決策樹迴歸可以捕捉非線性關係，同時也可以處理缺失資料，但容易過擬合。隨機森林迴歸在處理高維資料和大量資料時表現優秀，同時可以有效地避免過擬合。

在使用這些迴歸演算法時，需要注意一些問題。首先，過擬合是迴歸演算法中常見的問題。過擬合的主要原因是模型複雜度過高，而樣本數量不足。因此，在訓練模型時，需要使用交叉驗證等方法評估模型的泛化能力，並選擇合適的正則化方法來控制模型的複雜度。另外，在資料預處理時，需要注意特徵的標準化和歸一化，以及異常值的處理，這些都會影響模型的效果。

總的來說，每種演算法都有其適用的場景和優缺點，需要結合具體的應用場景進行選擇。在使用演算法時，還需要對資料進行預處理，包括缺失值填充、特徵縮放、特徵選擇等，以提高演算法的預測能力和魯棒性。
