# 7大經典回歸模型總結

3D視覺工程師 2022-03-14 21:27

今天給大家介紹機器學習建模中7大經典的回歸分析模型。

## 什麼是回歸分析？

回歸分析是一種預測性的建模技術，它研究的是因變量（目標）和自變量（預測器）之間的關係。這種技術通常用於預測分析，時間序列模型以及發現變量之間的因果關係。例如，司機的魯莽駕駛與道路交通事故數量之間的關係，最好的研究方法就是回歸。

回歸分析是建模和分析數據的重要工具。在這裡，我們使用曲線/線來擬合這些數據點，在這種方式下，從曲線或線到數據點的距離差異最小。我會在接下來的部分詳細解釋這一點。

## 我們為什麼使用回歸分析？

如上所述，回歸分析估計了兩個或多個變量之間的關係。下面，讓我們舉一個簡單的例子來理解它：

比如說，在當前的經濟條件下，你要估計一家公司的銷售額增長情況。現在，你有公司最新的數據，這些數據顯示出銷售額增長大約是經濟增長的2.5倍。那麼使用回歸分析，我們就可以根

據當前和過去的信息來預測未來公司的銷售情況。

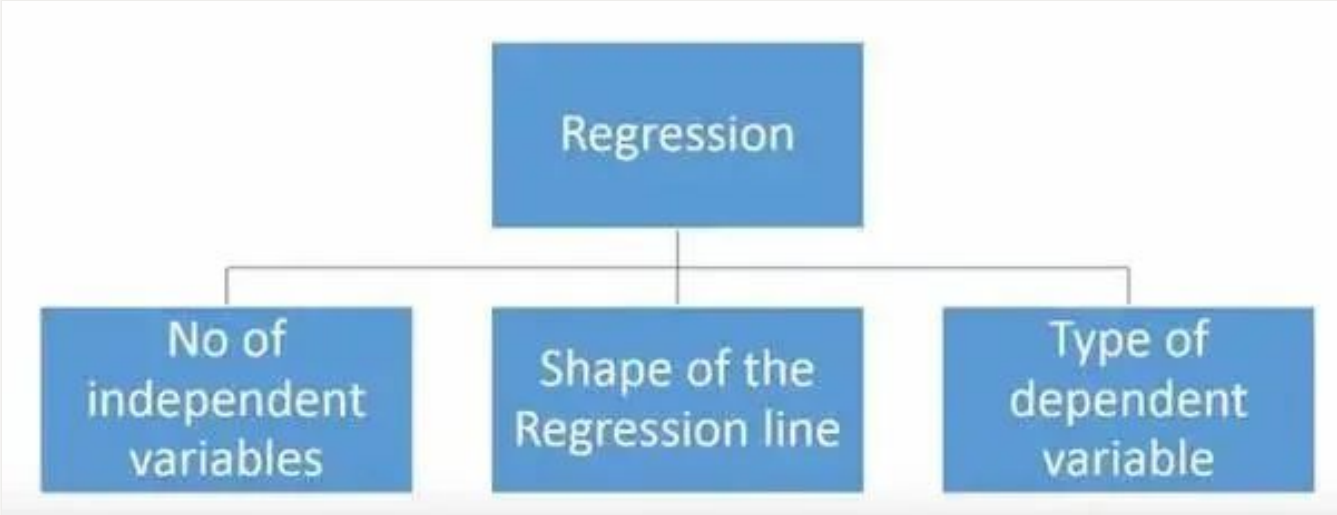
使用回歸分析的好處良多。具體如下：

1. 它表明自變量和因變量之間的顯著關係；
2. 它表明多個自變量對一個因變量的影響強度。

回歸分析也允許我們去比較那些衡量不同尺度的變量之間的相互影響，如價格變動與促銷活動數量之間聯繫。這些有利於幫助市場研究人員，數據分析人員以及數據科學家排除並估計出一組最佳的變量，用來構建預測模型。

**我們有多少種回歸技術？**

有各種各樣的回歸技術用於預測。這些技術主要有三個度量（自變量的個數，因變量的類型以及回歸線的形狀）。我們將在下面的部分詳細討論它們。



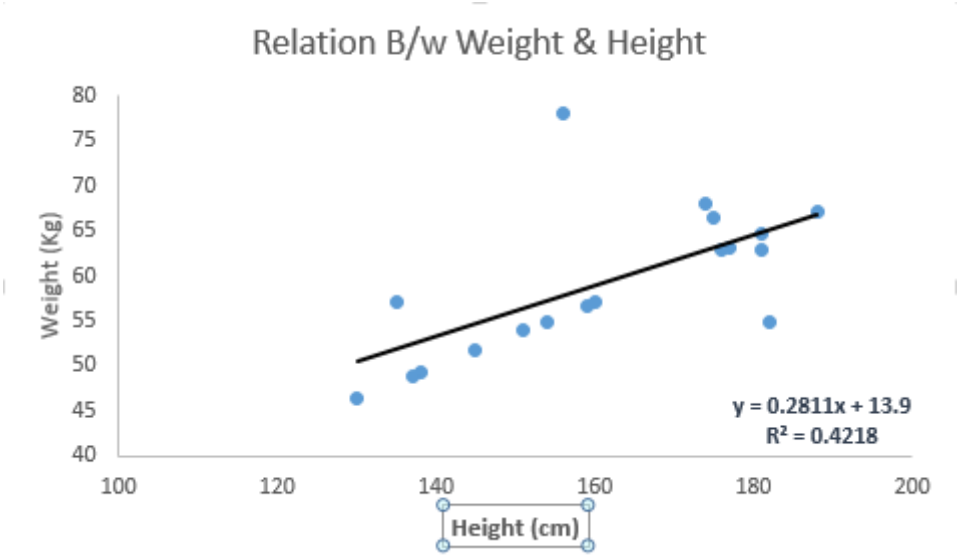

對於那些有創意的人，如果你覺得有必要使用上面這些參數的一個組合，你甚至可以創造出一個沒有被使用過的回歸模型。但在你開始之前，先了解如下最常用的回歸方法：

## 1.Linear Regression線性回歸

它是最為人熟知的建模技術之一。線性回歸通常是人們在學習預測模型時首選的技術之一。在這種技術中，因變量是連續的，自變量可以是連續的也可以是離散的，回歸線的性質是線性的。

線性回歸使用最佳的擬合直線（也就是回歸線）在因變量（Y）和一個或多個自變量（X）之間建立一種關係。

用一個方程式來表示它，即Y=a+b*X+e，其中a表示截距，b表示直線的斜率，e是誤差項。這個方程可以根據給定的預測變量（s）來預測目標變量的值。

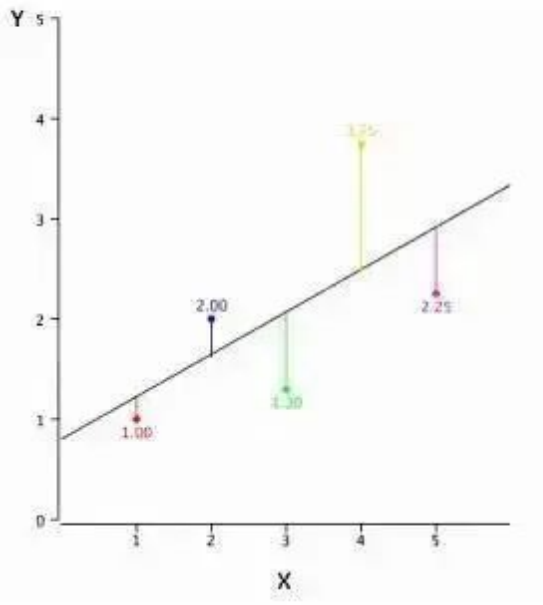



一元線性回歸和多元線性回歸的區別在於，多元線性回歸有（>1）個自變量，而一元線性回歸通常只有1個自變量。現在的問題是“我們如何得到一個最佳的擬合線呢？”。

**如何獲得最佳擬合線（a和b的值）？**

這個問題可以使用最小二乘法輕鬆地完成。最小二乘法也是用於擬合回歸線最常用的方法。對於觀測數據，它通過最小化每個數據點到線的垂直偏差平方和來計算最佳擬合線。因為在相加時，偏差先平方，所以正值和負值沒有抵消。

$$\min_{w} ||Xw - y||_2^2$$




我們可以使用R-square指標來評估模型性能。想了解這些指標的詳細信息，可以閱讀：模型性能指標Part 1,Part 2.

> **要點：**
>
> 1.自變量與因變量之間必須有線性關係
>
> 2.多元回歸存在多重共線性，自相關性和異方差性。
>
> 3.線性回歸對異常值非常敏感。它會嚴重影響回歸線，最終影響預測值。
>
> 4.多重共線性會增加係數估計值的方差，使得在模型輕微變化下，估計非常敏感。結果就是係數估計值不穩定
>
> 5.在多個自變量的情況下，我們可以使用向前選擇法，向後剔除法和逐步篩選法來選擇最重要的自變量。

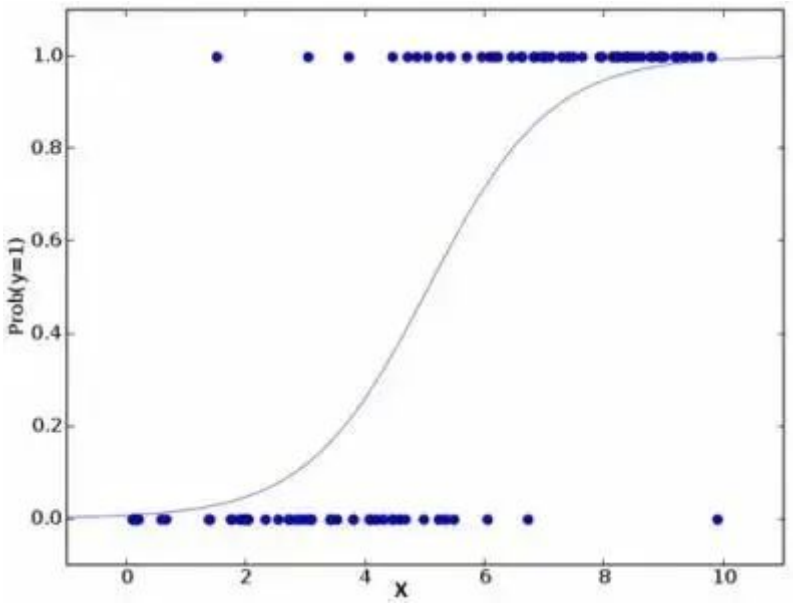
## 2.Logistic Regression邏輯回歸

邏輯回歸是用來計算“事件=Success”和“事件=Failure”的概率。當因變量的類型屬於二元（1 / 0，真/假，是/否）變量時，我們就應該使用邏輯回歸。這裡，Y的值從0到1，它可以用下方程表示。

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrenceln(odds) = ln(p/(1-p))logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

上述式子中，p表述具有某個特徵的概率。你應該會問這樣一個問題："我們為什麼要在公式中使用對數log呢？"。

因為在這裡我們使用的是的二項分佈（因變量），我們需要選擇一個對於這個分佈最佳的連結函數。它就是Logit函數。在上述方程中，通過觀測樣本的極大似然估計值來選擇參數，而不是最小化平方和誤差（如在普通回歸使用的）。




**要點：**

1.它廣泛的用於分類問題。

2.邏輯回歸不要求自變量和因變量是線性關係。它可以處理各種類型的關係，因為它對預測的相對風險指數OR使用了一個非線性的log轉換。

3.為了避免過擬合和欠擬合，我們應該包括所有重要的變量。有一個很好的方法來確保這種情況，就是使用逐步篩選方法來估計邏輯回歸。

4.它需要大的樣本量，因為在樣本數量較少的情況下，極大似然估計的效果比普通的最小二乘法差。

5.自變量不應該相互關聯的，即不具有多重共線性。然而，在分析和建模中，我們可以選擇包含分類變量相互作用的影響。

6.如果因變量的值是定序變量，則稱它為序邏輯回歸。

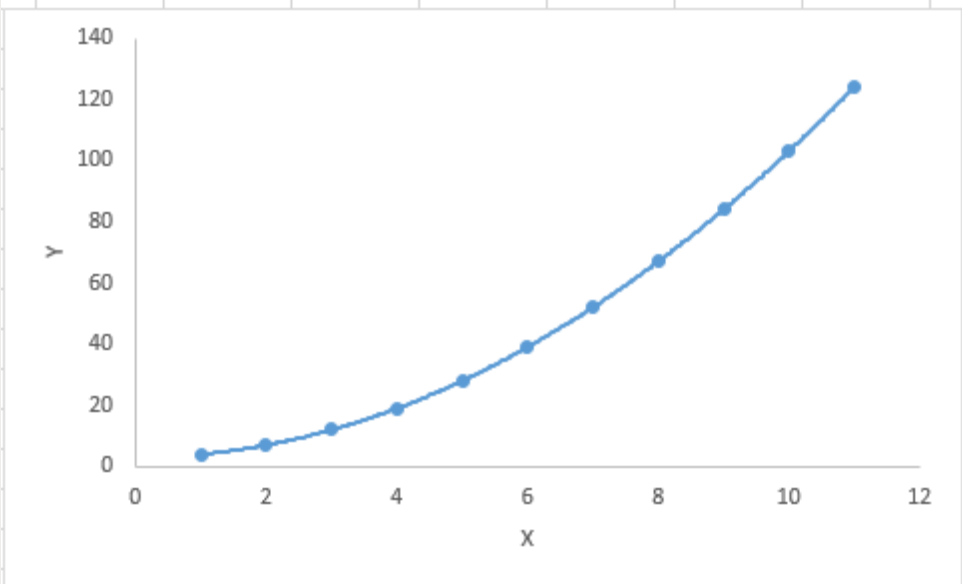
7.如果因變量是多類的話，則稱它為多元邏輯回歸。

## 3.Polynomial Regression多項式回歸

對於一個回歸方程，如果自變量的指數大於1，那麼它就是多項式回歸方程。如下方程所示：

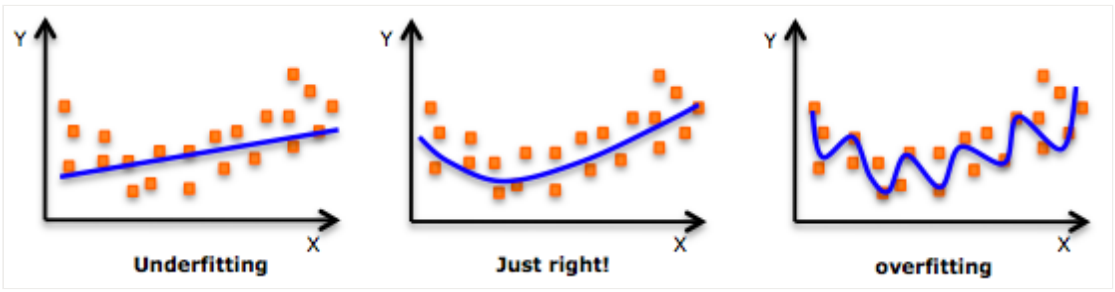
y=a+b*x^2

在這種回歸技術中，最佳擬合線不是直線。而是一個用於擬合數據點的曲線。




> 重點：
>
> 雖然會有一個誘導可以擬合一個高次多項式並得到較低的錯誤，但這可能會導致過擬合。你需要經常畫出關係圖來查看擬合情況，並且專注於保證擬合合理，既沒有過擬合又沒有欠擬合。

下面是一個圖例，可以幫助理解：



明顯地向兩端尋找曲線點，看看這些形狀和趨勢是否有意義。更高次的多項式最後可能產生怪異的推斷結果。

## 4.Stepwise Regression逐步回歸

在處理多個自變量時，我們可以使用這種形式的回歸。在這種技術中，自變量的選擇是在一個自動的過程中完成的，其中包括非人為操作。

這一壯舉是通過觀察統計的值，如R-square，t-stats和AIC指標，來識別重要的變量。逐步回歸通過同時添加/刪除基於指定標準的協變量來擬合模型。

下面列出了一些最常用的逐步回歸方法：

- 標準逐步回歸法做兩件事情。即增加和刪除每個步驟所需的預測。
- 向前選擇法從模型中最顯著的預測開始，然後為每一步添加變量。
- 向後剔除法與模型的所有預測同時開始，然後在每一步消除最小顯著性的變量。

這種建模技術的目的是使用最少的預測變量數來最大化預測能力。這也是處理高維數據集的方法之一。

## 5.Ridge Regression嶺回歸

嶺回歸分析是一種用於存在多重共線性（自變量高度相關）數據的技術。在多重共線性情況下，儘管最小二乘法（OLS）對每個變量很公平，但它們的差異很大，使得觀測值偏移並遠離真實值。嶺回歸通過給回歸估計上增加一個偏差度，來降低標準誤差。

上面，我們看到了線性回歸方程。還記得嗎？它可以表示為：

y=a+b*x這個方程也有一個誤差項。完整的方程是：

y=a+b*x+e (error term), [error term is the value needed to correct for a prediction error between the observed and predicted value]

=> y=a+y= a+ b1x1+ b2x2+....+e, for multiple independent variables.

在一個線性方程中，預測誤差可以分解為2個子分量。一個是偏差，一個是方差。預測錯誤可能會由這兩個分量或者這兩個中的任何一個造成。在這裡，我們將討論由方差所造成的有關誤差。

嶺回歸通過收縮參數λ（lambda）解決多重共線性問題。看下面的公式

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

在這個公式中，有兩個組成部分。第一個是最小二乘項，另一個是β2（β-平方）的λ倍，其中β是相關係數。為了收縮參數把它添加到最小二乘項中以得到一個非常低的方差。

> 要點：
>
> 1.除常數項以外，這種回歸的假設與最小二乘回歸類似；
>
> 2.它收縮了相關係數的值，但沒有達到零，這表明它沒有特徵選擇功能
>
> 3.這是一個正則化方法，並且使用的是L2正則化。

## 6.Lasso Regression套索回歸

它類似於嶺回歸，Lasso（Least Absolute Shrinkage and Selection Operator）也會懲罰回歸係數的絕對值大小。此外，它能夠減少變化程度並提高線性回歸模型的精度。看看下面的公式：

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

Lasso 回歸與Ridge回歸有一點不同，它使用的懲罰函數是絕對值，而不是平方。這導致懲罰（或等於約束估計的絕對值之和）值使一些參數估計結果等於零。使用懲罰值越大，進一步估

計會使得縮小值趨近於零。這將導致我們要從給定的n個變量中選擇變量。

> 要點：
>
> 1.除常數項以外，這種回歸的假設與最小二乘回歸類似；
>
> 2.它收縮係數接近零（等於零），這確實有助於特徵選擇；
>
> 3.這是一個正則化方法，使用的是L1正則化；
>
> 如果預測的一組變量是高度相關的，Lasso 會選出其中一個變量並且將其它的收縮為零。

## 7.ElasticNet回歸

ElasticNet是Lasso和Ridge回歸技術的混合體。它使用L1來訓練並且L2優先作為正則化矩陣。當有多個相關的特徵時，ElasticNet是很有用的。Lasso 會隨機挑選他們其中的一個，而ElasticNet則會選擇兩個。

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

Lasso和Ridge之間的實際的優點是，它允許ElasticNet繼承循環狀態下Ridge的一些穩定性。

> **要點：**
>
> 1.在高度相關變量的情況下，它會產生群體效應；
>
> 2.選擇變量的數目沒有限制；
>
> 3.它可以承受雙重收縮。

除了這7個最常用的回歸技術，你也可以看看其他模型，如Bayesian、Ecological和Robust回歸。

**如何正確選擇回歸模型？**

當你只知道一個或兩個技術時，生活往往很簡單。我知道的一個培訓機構告訴他們的學生，如果結果是連續的，就使用線性回歸。如果是二元的，就使用邏輯回歸！然而，在我們的處理中，可選擇的越多，選擇正確的一個就越難。類似的情況下也發生在回歸模型中。

在多類回歸模型中，基於自變量和因變量的類型，數據的維數以及數據的其它基本特徵的情況下，選擇最合適的技術非常重要。以下是你要選擇正確的回歸模型的**關鍵因素**：

1.數據探索是構建預測模型的必然組成部分。在選擇合適的模型時，比如識別變量的關係和影響時，它應該首選的一步。

2. 比較適合於不同模型的優點，我們可以分析不同的指標參數，如統計意義的參數，R-square，Adjusted R-square，AIC，BIC以及誤差項，另一個是Mallows' Cp準則。這個主要是通過將模型與所有可能的子模型進行對比（或謹慎選擇他們），檢查在你的模型中可能出現的偏差。

3.交叉驗證是評估預測模型最好額方法。在這裡，將你的數據集分成兩份（一份做訓練和一份做驗證）。使用觀測值和預測值之間的一個簡單均方差來衡量你的預測精度。

4.如果你的數據集是多個混合變量，那麼你就不應該選擇自動模型選擇方法，因為你應該不想在同一時間把所有變量放在同一個模型中。

5.它也將取決於你的目的。可能會出現這樣的情況，一個不太強大的模型與具有高度統計學意義的模型相比，更易於實現。

6.回歸正則化方法（Lasso，Ridge和ElasticNet）在高維和數據集變量之間多重共線性情況下運行良好。

作者：Sunil Ray（譯者:劉帝偉） 來源：csdn

閱讀原文

喜歡此內容的人還喜歡

NMI | 通過生成模型進行單一片段修飾的分子優化

人工智能藥物設計

機器學習必知必會10大算法!

深度學習初學者

牛頓迭代法的可視化詳解

深度學習初學者