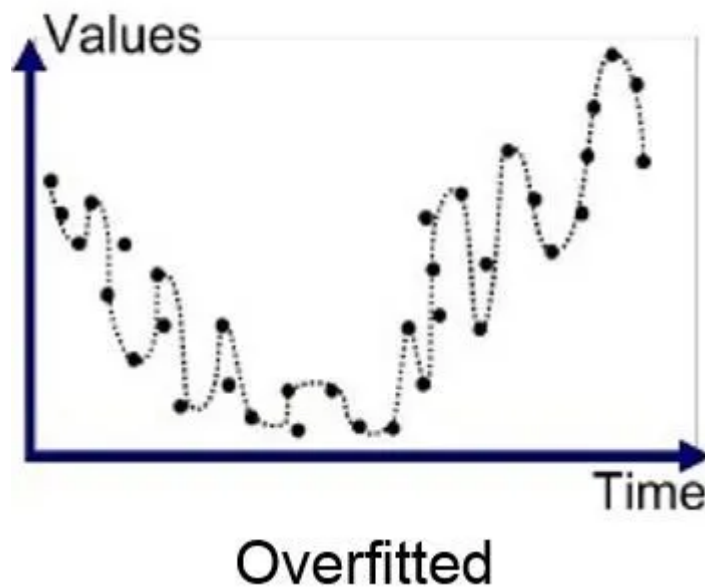


# 防止模型過擬合的必備方法！

小白 小白學視覺 昨天

點擊上方

重磅乾貨，第一時間送達



作者：Mahitha

來源：機器之心

正如巴菲特所言：

在機器學習中，過擬合（overfitting）會使模型的預測性能變差，通常發生在模型過於複雜的情況下，如參數過多等。本文對過擬合及其解決方法進行了歸納闡述。

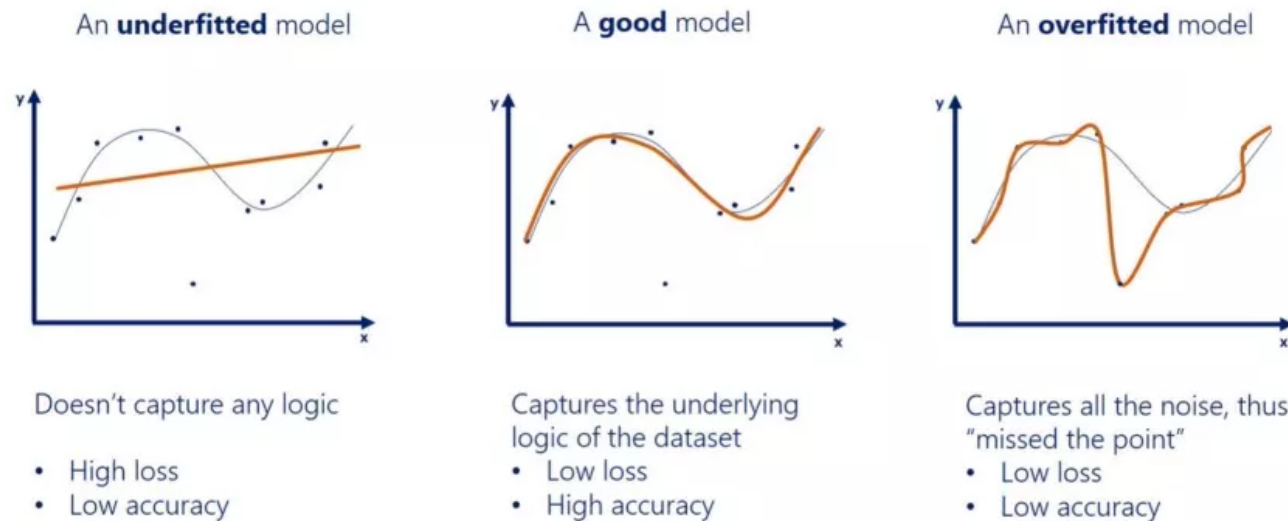
**Training  
Set**

**Validation  
Set**

**Test  
Set**



在機器學習中，如果模型過於專注於特定的訓練數據而錯過了要點，那麼該模型就被認為是過擬合。該模型提供的答案和正確答案相距甚遠，即準確率降低。這類模型將無關數據中的噪聲視為信號，對準確率造成負面影響。即使模型經過很好地訓練使損失很小，也無濟於事，它在新數據上的性能仍然很差。欠擬合是指模型未捕獲數據的邏輯。因此，欠擬合模型具備較低的準確率和較高的損失。



### 如何確定模型是否過擬合？

構建模型時，數據會被分為3類：訓練集、驗證集和測試集。訓練數據用來訓練模型；驗證集用於在每一步測試構建的模型；測試集用於最後評估模型。通常數據以80:10:10 或70:20:10 的比率分配。

在構建模型的過程中，在每個epoch 中使用驗證數據測試當前已構建的模型，得到模型的損失和準確率，以及每個epoch 的驗證損失和驗證準確率。模型構建完成後，使用測試數據對模型進行測試並得到準確率。如果準確率和驗證準確率存在較大的差異，則說明該模型是過擬合的。

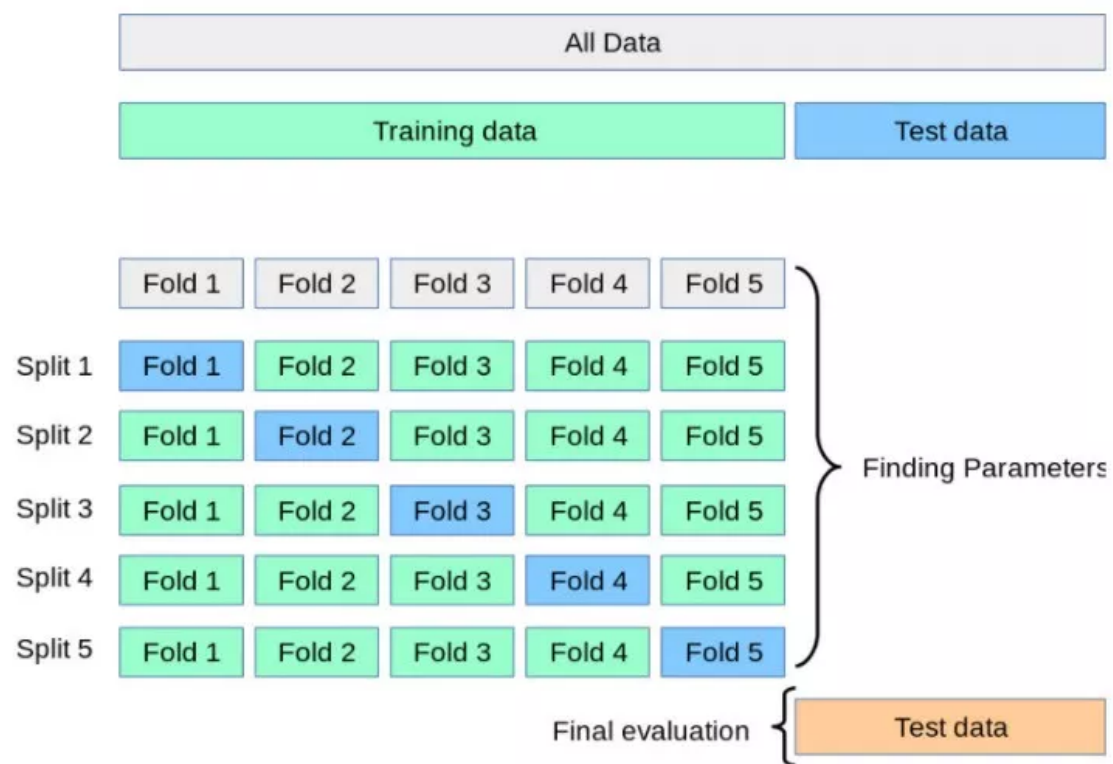
如果驗證集和測試集的損失都很高，那麼就說明該模型是欠擬合的。

### 如何防止過擬合

## 交叉驗證

交叉驗證是防止過擬合的好方法。在交叉驗證中，我們生成多個訓練測試劃分（splits）並調整模型。K-折驗證是一種標準的交叉驗證方法，即將數據分成k 個子集，用其中一個子集進行驗證，其他子集用於訓練算法。

交叉驗證允許調整超參數，性能是所有值的平均值。該方法計算成本較高，但不會浪費太多數據。交叉驗證過程參見下圖：



## 用更多數據進行訓練

用更多相關數據訓練模型有助於更好地識別信號，避免將噪聲作為信號。數據增強是增加訓練數據的一種方式，可以通過翻轉（flipping）、平移（translation）、旋轉（rotation）、縮放（scaling）、更改亮度（changing brightness）等方法來實現。

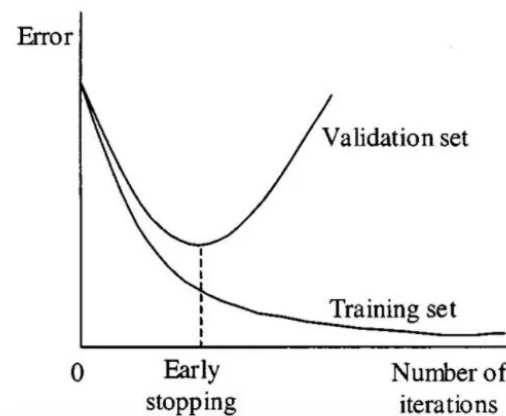
## 移除特徵

移除特徵能夠降低模型的複雜性，並且在一定程度上避免噪聲，使模型更高效。為了降低複雜度，我們可以移除層或減少神經元數量，使網絡變小。

## 早停

對模型進行迭代訓練時，我們可以度量每次迭代的性能。當驗證損失開始增加時，我們應該停止訓練模型，這樣就能阻止過擬合。

下圖展示了停止訓練模型的時機：



## 正則化

正則化可用於降低模型的複雜性。這是通過懲罰損失函數完成的，可通過L1 和L2 兩種方式完成，數學方程式如下：

$$L(x, y) \equiv \sum_{i=1}^n (y_i - h_{\theta}(x_i))^2 + \lambda \sum_{i=1}^n |\theta_i|$$

L1 懲罰的目的是優化權重絕對值的總和。它生成一個簡單且可解釋的模型，且對於異常值是魯棒的。

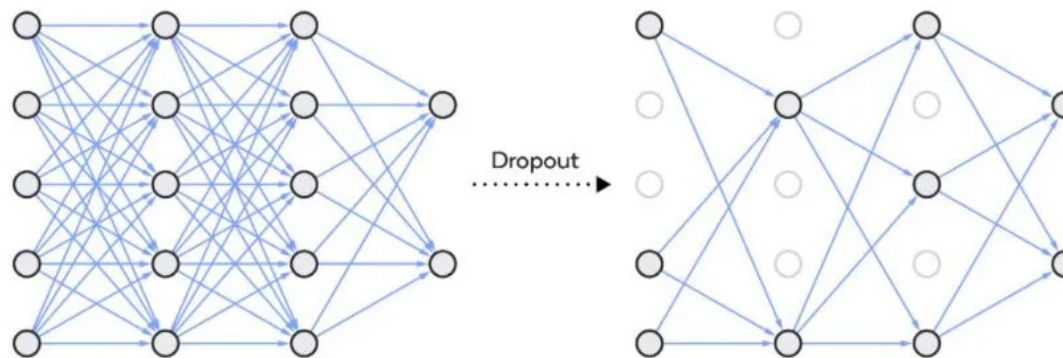
$$L(x, y) \equiv \sum_{i=1}^n (y_i - h_{\theta}(x_i))^2 + \lambda \sum_{i=1}^n \theta_i^2$$

L2 懲罰權重值的平方和。該模型能夠學習複雜的數據模式，但對於異常值不具備魯棒性。

這兩種正則化方法都有助於解決過擬合問題，讀者可以根據需要選擇使用。

## Dropout

Dropout 是一種正則化方法，用於隨機禁用神經網絡單元。它可以在任何隱藏層或輸入層上實現，但不能在輸出層上實現。該方法可以免除對其他神經元的依賴，進而使網絡學習獨立的相關性。該方法能夠降低網絡的密度，如下圖所示：



### 總結

過擬合是一個需要解決的問題，因為它會讓我們無法有效地使用現有數據。有時我們也可以在構建模型之前，預估到會出現過擬合的情況。通過查看數據、收集數據的方式、採樣方式，錯誤的假設，錯誤表徵能夠發現過擬合的預兆。為避免這種情況，請在建模之前先檢查數據。但有時在預處理過程中無法檢測到過擬合，而是在構建模型後才能檢測出來。我們可以使用上述方法解決過擬合問題。

### 下載1: OpenCV-Contrib擴展模塊中文版教程

在「

### 下載2: Python視覺實戰項目52講

在

### 下載3: OpenCV實戰項目20講

在

### 交流群

歡迎加入公眾號讀者群一起和同行交流，目前有 請按照格式備註，否則不予通過 添加成功後會根據研究方向邀請進入相關微信群。請勿



A promotional banner for '小白学视觉' (Xiao Bai Xue Shi Jue). The background is black with a white piano keyboard on the right. On the left, there is a blue circular logo with a white robot icon. The text '小白学视觉' is in large white characters. Below it, '计算机视觉' (Computer Vision) is written in white. Further down, '论文解读 求职感想' (Paper interpretation, job-seeking thoughts) and 'SLAM技术 深度学习 学习感受' (SLAM technology, deep learning, learning experience) are listed in white. A dashed line separates this from the bottom text: '距离我们只差一个' (Only one step away from us) in white, and '长按关注' (Long press to follow) in red. On the right side, the text '聚集地' (Gathering place) and '计算机视觉学者' (Computer vision scholars) is written vertically in white. At the bottom right, there is a QR code with a blue border and a small robot icon in the center.

喜歡此內容的人還喜歡



## 聊一聊“超大模型”

磐創AI



## 遙遙無期還是近在咫尺？長文展望大模型商業化前景

學術頭條



## DSOD回顧：從頭訓練深度監督目標檢測模型

機器學習研究組訂閱

