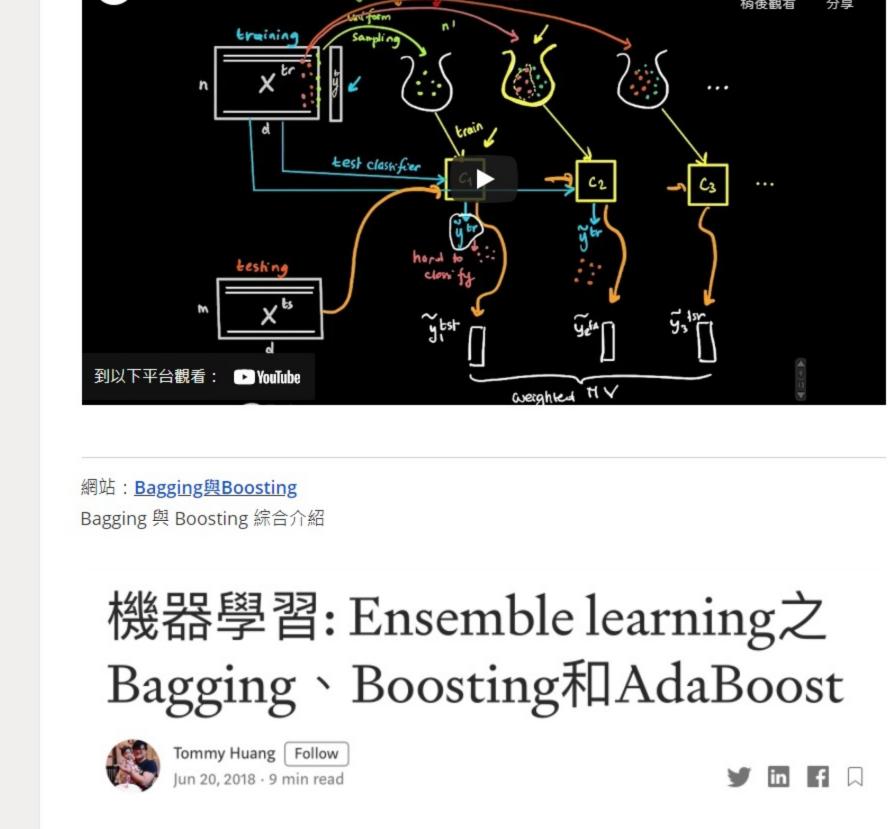
anton AI共學社群 我的 AI共學社群 > Part1 - NLP 經典機器學習馬拉松 > D26: Adaboost 演算法 D26: Adaboost 演算法 囯 =~ 學習心得(完成) 課程閱讀 本日作業 問題討論 重要知識點 集成學習(Ensemble NLP自然語言學習實戰馬拉松 Learning) ▶ Adaboost 演算法 **Boosting** NLP 陪跑專家: Leo Liou 劉冠宏 重要知識點 • 了解何謂集成 (Ensemble) 中的 Boosting • 了解何謂 Adaboost 集成學習(Ensemble Learning) 上堂課程在介紹隨機森林前,有初步介紹了 Ensemble Learning 的概念,這邊在跟同學做個詳細的介 紹。 一般的 Supervised Learning: Train Prediction Model Test • 當只有單一模型預測時,容易陷入 variance-bias trade-off 的問題中 • 單一模型陷入局部低點(local optimal) 時,模型無法獲的最佳結果 Ensemble • 透過集結不同的模型,增加複雜度,減少 variance • 透過平均或投票,集結眾多模型的預測,減少單一模型的盲點 集成學習有個重要的論點在於,所有用來集成的模型必須要夠歧異(Diverse)。 模型沒有歧異性,無法分出新的資料(粉紅色方框) 2 feature Model1 Model2 feature 1 模型有歧異性,集成後可分出新資料(粉紅色方框) feature 2 Model1 Model2 feature 1 Ensemble 中有不同的集成手法,最常見的有像是上個章節介紹過的 Bagging,也是 Random Forest 採用的集成方法。 並聯 bootstrap Training subset 1 (n) Classifier I Training subset 2 (n) Classifier 2 Training Set (N) Majority Vote Training subset L (n) Classifier L 由訓練集抽樣的 個別獨立模型 子資料集 而這章節主要要介紹的 Adaboost,則是另外一個常見的集成手法 Boosting。Boosting 是一種 Bagging 的變形,其背後的思想在於"Boost 模型的表現"。將模型以序列的方式串接再一起,透過加 強學習前一步的錯誤,來增強這一步模型的表現 第一個分類器 Initial weight: Weight 1 for Training set Classifier 1 Result 1 equivalent Weight 2 for Training set Result 2 Weighted-sum 第二個分類器 ◆ Classifier L Weight L for Training set Result I. 第L個分類器 ◆ 串聯(序列) **Boosting** 透過對訓練集進行取樣 (re-sample) 得到不同獨立的子訓練集 (subset),利用這些子集對模型進行訓 練,已取得個別不同的模型來進行 Ensemble。Boosting 也是使用這類的方式來進行訓練,與 Bagging 不同的是 Boosting 在取樣是會著重在分錯的樣本上。 接下來讓我們來看看 Boosting 演算法的步驟。 • 從訓練集均勻抽樣 (uniform sampling) 取出訓練子集 • 利用抽取的訓練子集訓練分類器 A • 利用訓練集資料測試訓練好的分類器 A • 取得分類錯誤的樣本 訓練集資料 訓練子集 1. 均匀抽樣 X Y 2. 訓練 3. 測試 4. 取得分類錯誤樣本 Classifier • 給予上輪分類錯誤的樣本較高的機率權重 (更容易被抽中) • 從訓練集非均勻抽樣 (non-uniform sampling) 取出訓練子集 (分錯的樣本較易被抽中) • 利用抽取的訓練子集訓練分類器 B • 利用訓練集資料測試訓練好的分類器 B • 取得分類錯誤的樣本 較高權重 訓練集資料 訓練子集 5.非均匀抽樣 Y X 6. 訓練 7. 測試 8. 取得分類錯誤樣本 Classifier • 給予上輪分類錯誤的樣本較高的機率權重 (更容易被抽中) • 從訓練集非均勻抽樣 (non-uniform sampling) 取出訓練子集 (分錯的樣本較易被抽中) 利用抽取的訓練子集訓練分類器 C • 利用訓練集資料測試訓練好的分類器 C • 取得分類錯誤的樣本 較高權重 訓練集資料 訓練子集 9.非均匀抽樣 X Υ 10. 訓練 12. 取得分類錯誤樣本 11. 測試 Classifier C 重複上述步驟取的所需的N個分類器 Classifier Classifier Classifier Classifier C В 再進行預測時,Bagging 使採用均勻多數決 (majority voting),而 Boosting 雖然也是採用多數決,但 Boosting 會依據分類器在使用訓練集做測試時表現給予不同權重,大家都有投票權·但每張票的價值 不同。在訓練時準確度較高的給較高的權重,相反的準確度較低的給予較低的權重。 Testing data Classifier Classifier Classifier Classifier W3Y3 WNYN 給予不同權重 Adaboost Adaboost 是一種對 Boosting 做改良的演算法,其在訓練每個模型時不重新取樣訓練資料,而是透過 將**分類錯誤的樣本權重提高**,來使每次新訓練的模型都聚焦在容易分類錯誤的樣本上。再進行預測時 一樣採用加權投票機制。Adaboost 主要的算法概念如下: • 使用訓練資料訓練第一個分類器,且此分類器錯誤率小於 0.5 • 改變訓練資料集的權重 (re-weighting training data), 使前一次訓練的分類器錯誤率等於 0.5 • 使用改變權重的訓練資料集訓練新的分類器 • 重複上述步驟,直到取的需要的分類器數量 • 預測時根據訓練時的錯誤率給予相對應的權重 (錯誤率高 --> 權重低) 文字看不懂怎麼辦? 讓我們以圖是來說明 Adaboost 的訓練流程 改變權重使分類器_1 一共五個樣本 分類器_1 結果 的錯誤率=0.5 0 分類器_2 0 0 錯誤率 = 1/5 = 0.2 新錯誤率 樣本權重 (初始為1) (第三個樣本分錯) 2/(1/2*4+2)=0.5Adaboost 優缺點 了解 Adaboost 的運作後,來看看 Adaboost 的特性與優缺點 優點: • 良好的利用弱分類器進行集成 • 具有高的精度 • 充分考慮不同分類器的權重分佈 缺點: • 數據不平衡時會使分類精度下降 • 對資料的噪聲較為敏感 • 訓練所需的時間較長 **Bagging V.S. Boosting Bagging Boosting** 使用均匀採樣 非均勻採樣 (樣本被取得的機率相等) (分類錯誤樣本被取得機率較高) 每輪訓練資料使用採樣的方式進行 每輪訓練資料使用權重分配的方式 模型為平行架構 模型為序列架構 使用均匀多數決 使用加權多數決 知識點回顧 在這章節我們學習到了 • 了解集成 (Ensemble) 中的 Boosting 了解 Adaboost 模型與其運作原理 延伸閱讀 網站: Boosting介紹 Bagging 介紹 YouTube 影片 Machine Learning 13.3 (Ensemble Learning: Boosting and AdaBoost) test classifier



Bagging、Boosting和AdaBoost (Adaptive Boosting)都是Ensemble

麼為什麼不用多個分類器呢?

В

learning(集成學習)的方法(手法)。Ensemble learning在我念書的時後我

比較喜歡稱為多重辨識器,名稱很直覺,就是有很多個辨識器。其概念

就是「三個臭皮匠勝過一個諸葛亮」,如果單個分類器表現的很好,那

一個分類器/hyperplane分不出來 兩個分類器/hyperplane就分的出來 Ensemble Learning基本條件是:每個分類器之間應該要有差異,每個分類器準確率需大於0.5。 如果用的分類器沒有差異,那只是用很多個一樣的分類器來分類,結果 合成起來是沒有差異的。如果分類器的精度p<0.5,隨著ensemble規模的

В

A