07/22 DM 上午場

洪子軒

Sent: Friday, July 22, 2016 1:38 PM
To: 洪子軒

【時間序列關連】

找出經常循序出現的項目組合，進而瞭解顧客的長期行為

＞排序
＞轉換：用 Large Itemset 取代原記錄

以每個顧客為單位來衡量，看他們是否具有相似性
ex 買完xbox 機後 → 會再來買 VR 眼鏡

間隔時間

--
【分類】

Overfitting：過了一個點，測試資料誤差上升
也要考慮成本議題，分類器太複雜（noise point 還是會造成干擾）

Underfitting：例如欄位太少，沒有代表性
→四個方法的檢定（相關係數、ANAOVA、T-檢定…）

Holdout：部份（50%以上）建模、部份預測
Cross validation：sliding window 的方式 k-fold （通常切五段）
→變異數小，建模正確率佳

＞KNN（近鄰法）：快、易得結果（k 一般取奇數，輪流測 1 3 5 7 9）
圖上會出現兩個點的垂直平分線
k 太小會受到雜訊的干擾，太大鄰居

結果：confusion matrix 混淆矩陣

--
洪子軒 Tzu-Hsuan Hung
中華電信研究院 巨量資料所
TEL: (03)-4245128
Email: Lucas@cht.com.tw
32661桃園市楊梅區電研路99號
--