07/25 DM 上午場

洪子軒

Sent: Monday, July 25, 2016 11:56 AM
To: 洪子軒

【分類】

http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

＞Logistics 迴歸：可以作為 base line，簡單線性函數
X 自變數：可以是類別/連續、常態分配為佳，但也可以是非常態分配
Y 應變數：類別
e exp 次方
ln 自然對數

p= e^f(X) / (1 + e^f(X))
f(X) = ln(P / (1-P)) = β0 + β1X1 + β2X2 + ...

結果：機率（比類別標籤細膩）

※Female/Mail 英文字母順序的關係，常見編碼為 0/1

＞Probit 分類（類似累積機率密度函數）
需要常態分配

http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

＞決策樹

http://scikit-learn.org/stable/modules/tree.html

・當 X 給得很多時，容易過度學習
・純度 Homogeneous EX 9比1
・Gini（CART 演算法）亂度或不純度
・Entropy（C4.5）
・屬性會被選上是因為它有很高的分類貢獻
・數值分割臨界值
・Gain Ratio：為了懲罰（penalized）分很多段的屬性，讓它不容易被選到
→例如：員工代號

scikit-learn uses an optimised version of the CART algorithm

--
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
--