

第三屆商業模式與大數據分析競賽 人工智慧金融挑戰賽

R語言資料科學應用工作坊範例教學

2020/09/26

蘇彥庭

分類問題

範例資料集簡介

- 資料集名稱：Bank Marketing Data Set
- 資料來源：[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
- 此資料為葡萄牙某家銀行的電話行銷活動數據，同一個客戶有可能被多次行銷，並記錄是否會申辦定期存款。
- 資料集預測目標：預測客戶是否會申辦定期存款(分類問題)

範例資料集下載

- <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#>



Bank Marketing Data Set

Download [Data Folder](#) [Data Set Description](#)

點選

Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).


Data Set Characteristics:	Multivariate	Number of Instances:	45211	Area:	Business
Attribute Characteristics:	Real	Number of Attributes:	17	Date Donated	2012-02-14
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	1277772

Source:

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

範例資料集下載

Index of /ml/machine-learning-databases/00222

- [Parent Directory](#)
 - [bank-additional.zip](#)
 - [bank.zip](#)
- 
- 點選下載

Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips SVN/1.7.14 Phusion_Passenger/4.0.53 mod_perl/2.0.11 Perl/

範例資料集內容

- bank.csv：4,521筆樣本，17個欄位，此範例將做為測試集資料
- bank-full.csv：45,211筆樣本，17個欄位，此範例將做為訓練集資料
- Bank-names.txt：資料說明



範例資料集變數說明

- 輸入變數：

1. age：年齡
2. job：職業
3. marital：婚姻狀況
4. education：教育程度
5. default：是否信用違約
6. balance：平均每年帳戶餘額(歐元)
7. housing：是否有房貸
8. loan：是否有個人貸款
9. contact：接觸類型(電話 or 手機)
10. day：最近接觸的日

範例資料集變數說明

- 輸入變數(承上頁)

11. month : 最近接觸的月

12. duration : 最近一次通話的長度(單位 : 秒)

13. campaign : 行銷活動期間內聯絡客戶的次數

14. pdays : 上一次行銷活動期間內與客戶最後一次聯絡後經過之天數
(-1表示上一次行銷活動沒有聯絡過)

15. previous : 本次行銷活動期間開始前和客戶聯絡的次數

16. poutcome : 上一次行銷活動成果(未知、其他、成功或失敗)

- 輸出變數

17. y : 客戶是否有申辦定期存款(成功或失敗)

混淆矩陣(Confusion Matrix)

$$\begin{cases} H_0 : \text{客戶不會申購定期存款} \\ H_1 : \text{客戶會申購定期存款} \end{cases}$$

- 型一錯誤：虛無假說事實上成立，統計檢驗的結果拒絕虛無假說
- 型二錯誤：虛無假說事實上不成立，但統計檢驗的結果不拒絕虛無假說

		實際	
		Positive	Negative
預測	Positive	TP (True Positive)	FP (False Positive) Type I Error
	Negative	FN (False Negative) Type II Error	TN (True Negative)

混淆矩陣(Confusion Matrix)

準確度(Accuracy) = $(TP + TN) / (TP + TN + FP + FN)$

精確度(Precision) = $TP / (TP + FP)$

召回率(Recall) = $TP / (TP + FN)$

F1 Score = $\frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 * \frac{Precision * Recall}{Precision + Recall}$

		實際	
		Positive	Negative
預測	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

迴歸問題

範例資料集簡介

- 資料集名稱：Real estate valuation data set Data Set
- 資料來源：

Original Owner and Donor

Name: Prof. I-Cheng Yeh

Institutions: Department of Civil Engineering, Tamkang University, Taiwan.

- 此資料為台灣新北市的房地產資料。
- 資料集預測目標：預測單位面積房價(新台幣1萬/坪)(迴歸問題)

範例資料集下載

- <https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set>



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)

☒ Repository ☐ Web 

[View ALL Data Sets](#)

Real estate valuation data set Data Set

Download: [Data Folder](#) [Data Set Description](#)

點選

Abstract: The “real estate valuation” is a regression problem. The market historical data set of real estate valuation are collected from Sindian Dist., New Taipei City, Taiwan.

Data Set Characteristics:	Multivariate	Number of Instances:	414	Area:	Business
Attribute Characteristics:	Integer, Real	Number of Attributes:	7	Date Donated	2018-08-18
Associated Tasks:	Regression	Missing Values?	N/A	Number of Web Hits:	78416

範例資料集下載

Index of /ml/machine-learning-databases/00477

- [Parent Directory](#)
- [Real estate valuation data set.xlsx](#)

點選下載

Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips SVN/1.7.14 Phusion_Passenger/4.0.53 mod_perl/2.0.11 Perl/v

範例資料集內容

- Real estate valuation data set.xlsx
- 共414筆樣本，7個欄位資料
- 輸入變數：
 - X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
 - X2=the house age (unit: year)
 - X3=the distance to the nearest MRT station (unit: meter)
 - X4=the number of convenience stores in the living circle on foot (integer)
 - X5=the geographic coordinate, latitude. (unit: degree) 緯度
 - X6=the geographic coordinate, longitude. (unit: degree) 經度
- 輸出變數：
 - Y= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

推薦學習資源

推薦學習資源

- 台大電機李宏毅副教授YouTube頻道
- https://www.youtube.com/watch?v=CXgbekl66jc&list=PLJV_el3uVTsPy9oCRY30oBPNLCo89yu49
- 台大資工林軒田教授YouTube頻道
- 機器學習基石
- <https://www.youtube.com/watch?v=nQvpFSMPhr0&list=PLXVfgk9fNX2I7tB6oIINGBmW50rrmFTqf>
- 機器學習技法
- <https://www.youtube.com/watch?v=A-GxGCCAArg&list=PLXVfgk9fNX2IQOYPmqjqWsNUFI2kpk1U2>

推薦學習資源

- XGBoost官網教學
- <https://xgboost.readthedocs.io/en/latest/R-package/index.html>

XGBoost

TABLE OF CONTENTS

- Installation Guide
- Get Started with XGBoost
- XGBoost Tutorials
- Frequently Asked Questions
- XGBoost User Forum
- GPU support
- XGBoost Parameters
- Python package
- R package**
 - Introduction to XGBoost in R
 - Understanding your dataset with XGBoost
- JVM package
- Ruby package

XGBoost R Package

CRAN **1.2.0.1** downloads **69K/month**

You have found the XGBoost R Package!

Get Started

- Checkout the [Installation Guide](#) contains instructions to install and how to use XGBoost for various tasks.
- Read the [API documentation](#).
- Please visit [Walk-through Examples](#).

Tutorials

- [Introduction to XGBoost in R](#)
- [Understanding your dataset with XGBoost](#)