

[資料分析&機器學習] 第3.3講：線性分類-邏輯斯回歸(Logistic Regression) 介紹



Yeh James · Follow

Published in JamesLearningNote

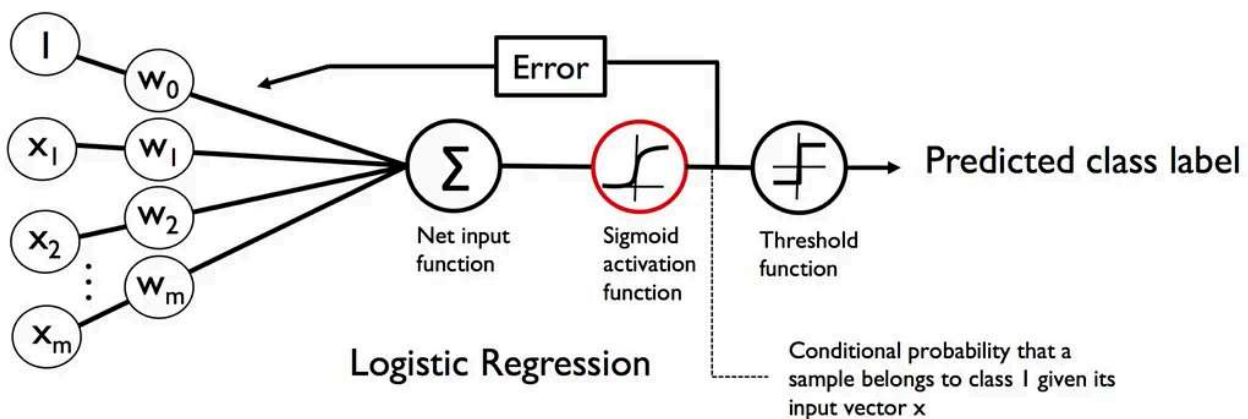
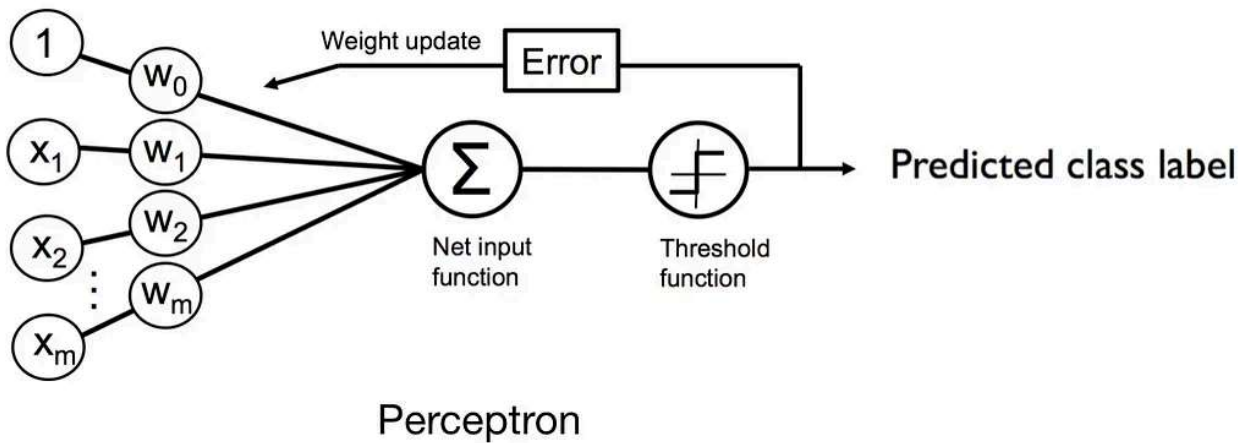
6 min read · Nov 3, 2017



Share

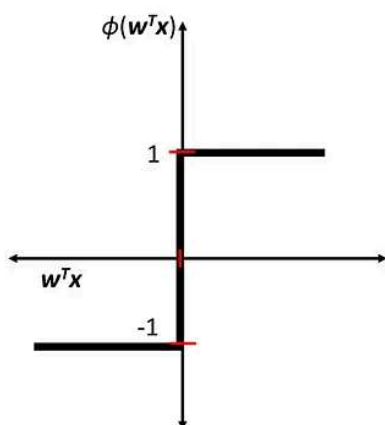
前面Perceptron 能夠讓我們成功達成二元分類，但我們只能知道預測結果是A還是B，沒辦法知道是A、是B的機率是多少。這種應用在我們生活中非常常見，比如說我們要根據今天的溫度、濕度、風向來預測明天的天氣，通常我們會需要知道明天是晴天的機率以及雨天的機率，來決定是否帶傘具出門。如果使用Logistic Regression就可以幫我們達成這樣的目標!

很重要的一點是Logistic Regression(邏輯斯回歸)很多人看名字以為是回歸的模型，但其實是一個分類的模型，名字取的不好很容易讓人誤解 XD。這個分類的模型大致跟Perceptron類似，只是Perceptron是根據 $w_0 \cdot x_0 + w_1 \cdot x_1 + \dots + w_n \cdot x_n > 0$ 或 ≤ 0 來判斷成A或B類，而Logistic Regression則是一個平滑的曲線，當 $w_0 \cdot x_0 + w_1 \cdot x_1 + \dots + w_n \cdot x_n$ 越大時判斷成A類的機率越大，越小時判斷成A類的機率越小。由於是二元分類，如果判斷成A類的機率越小，B類的機率越大(判斷成B類的機率 = 1 - 判斷成A的機率)。

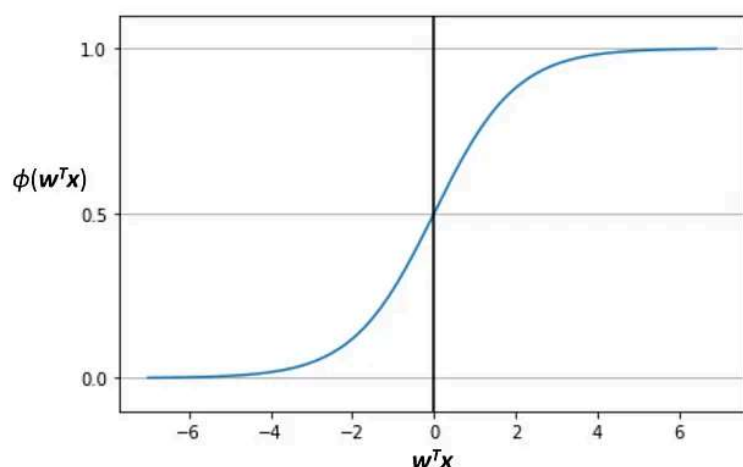


Perceptron以及Logistic Regression模型的差異

Perceptron



Logistic Regression



Perceptron、Logistic Regression激勵函數

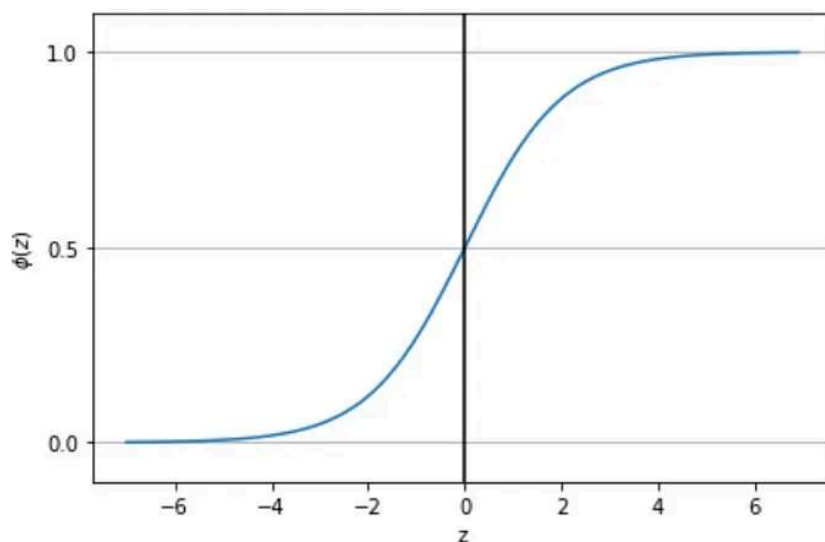
首先先介紹一下Sigmoid函數，也稱為logistic function，這個函數的 y 的值介於 $0 \sim 1$ ，這樣的分布也符合機率是在 $0 \sim 1$ 的範圍中。或許有人會覺得疑惑，Logistic Regression

為什麼要用這個Logistic函數？其實也可以改用其他符合0~1的函數（因為機率的值是介於0~1），只是Logistic函數是這種介於0~1的平滑函數中相對簡單的。

依下圖所示，當 $z=0$ 時判斷成+1類(A類)的機率為0.5，因此只要 $z > 0$ 判斷成A類的機率就會 >0.5 ，我們也就把它判斷成+1類(A類)。(這邊跟上一章perceptron一樣，只是多了機率的資訊) 如果 $z \leq 0$ 判斷成A類的機率就 ≤ 0.5 ，因此我們就把他判斷成-1類(B類)

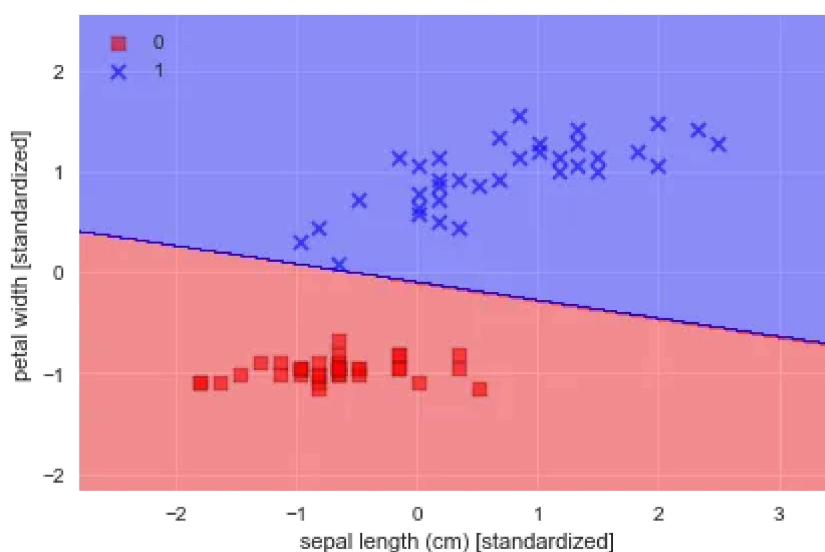
$$z = w^T x$$

$$\phi(z) = \frac{1}{1 + e^{-z}}$$



Logistic函數

接下來要說明要這個Logistic Regression要怎麼找到一條線，將兩群做線性分類，最終的結果如下圖所示



Logistic Regression不需要像上一個Perceptron演算法需要去看一個一個的資料點來做更新，Logistic Regression有一個數學解的方法可以直接找到一組 W ！

為了數學推導方便，之前我們將二元分類的A類以+1表示、B類以-1表示，現在將A類改以+1表示、B類以0表示。我們想要找到一組 w ，能夠將下方的式子變成最大值，那組 w 就是我們要找的線($z=w*x$)。下方的式子是希望當 $y=1$ 的時候 $\phi(z)$ 越靠近1 (判斷成A類的機率越大)，由於 $1-y$ 是0所以右邊的項會是1，當 $y=0$ 時左邊這項會是1右邊這項希望 $\phi(z)$ 越靠近0越好 (判斷成B類的機率越大)。

$$\prod_{i=1}^n \left(\phi(z^{(i)}) \right)^{y^{(i)}} \left(1 - \phi(z^{(i)}) \right)^{1-y^{(i)}}$$

我們可以使用微積分以及梯度下降的知識來讓上方的式子變為一個相對的最大值，有興趣的朋友可以參考Python機器學習這本書或是上Coursera參考吳恩達的機器學習課程。

接下來要教大家怎麼使用直接套用Sklearn裡面的logistic Regression model

載入Iris資料集

```
from sklearn import datasets
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
%matplotlib inline
```

```
iris = datasets.load_iris()
x = pd.DataFrame(iris['data'], columns=iris['feature_names'])
print("target_names: "+str(iris['target_names']))
y = pd.DataFrame(iris['target'], columns=['target'])
iris_data = pd.concat([x,y], axis=1)
iris_data = iris_data[['sepal length (cm)', 'petal length (cm)', 'target']]
iris_data = iris_data[iris_data['target'].isin([0,1])]
iris_data.head(3)
```

```
target_names: ['setosa' 'versicolor' 'virginica']
```

使用sklearn中的model_selection函式將把資料分為兩群train、test，將來可使用test資料來檢驗我們的分類模型效果

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(
    iris_data[['sepal length (cm)', 'petal length (cm)']], iris_data[['target']], test_size=0.3, random_state=0)
```

使用Logistic Regression之前需要先對資料做特徵縮放

```
from sklearn.preprocessing import StandardScaler
```

```
sc = StandardScaler()
sc.fit(X_train)
X_train_std = sc.transform(X_train)
X_test_std = sc.transform(X_test)
```

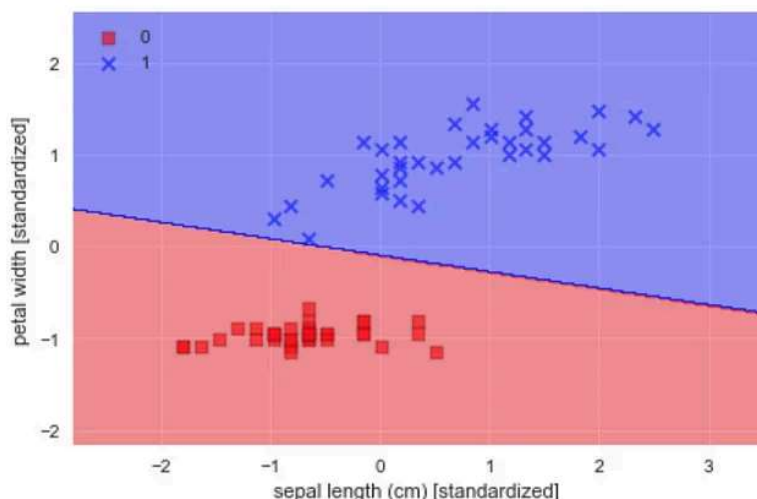
初始化Logistic Regression函式，以及將資料放進Logistic Regression開始訓練

```
lr = LogisticRegression()
lr.fit(X_train_std, y_train['target'].values)
```

```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
    intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
    penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
    verbose=0, warm_start=False)
```

視覺化訓練後的結果，可以明顯看出最後產出一條線將資料分為兩類

```
plot_decision_regions(X_train_std, y_train['target'].values, classifier=lr)
plt.xlabel('sepal length (cm) [standardized]')
plt.ylabel('petal width [standardized]')
plt.legend(loc='upper left')
plt.tight_layout()
plt.show()
```



預測test的資料看正確率多少？發現正確率100%!完美分類

```
lr.predict(X_test_std)
```

```
array([0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0])
```

Open in app ↗

Sign up

Sign in

Medium

Search



```
error = 0
for i, v in enumerate(lr.predict(X_test_std)):
    if v != y_test['target'].values[i]:
        error+=1
print(error)
```

0

使用Predict_proba函式，可知道預測的機率為多少

```
lr.predict_proba(X_test_std)
```

```
array([[ 0.93978177,  0.06021823],
       [ 0.005938  ,  0.994062  ],
       [ 0.97412756,  0.02587244],
       [ 0.0212674  ,  0.9787326  ],
       [ 0.0119407  ,  0.9880593  ],
       [ 0.32159479,  0.67840521],
       [ 0.95312111,  0.04687889],
       [ 0.0100283  ,  0.9899717  ],
       [ 0.00867294,  0.99132706],
       [ 0.03869904,  0.96130096],
       [ 0.06752495,  0.93247505],
       [ 0.05267873,  0.94732127],
       [ 0.01641248,  0.98358752],
       [ 0.98776985,  0.01223015],
       [ 0.95041495,  0.04958505],
       [ 0.94900694,  0.05099306],
       [ 0.98704203,  0.01295797],
       [ 0.9094647  ,  0.0905353  ],
       [ 0.93811592,  0.06188408],
       [ 0.97556281,  0.02443719],
       [ 0.93978177,  0.06021823],
       [ 0.04455382,  0.95544618],
       [ 0.96461079,  0.03538921],
       [ 0.04095329,  0.95904671],
       [ 0.96560356,  0.03439644],
```

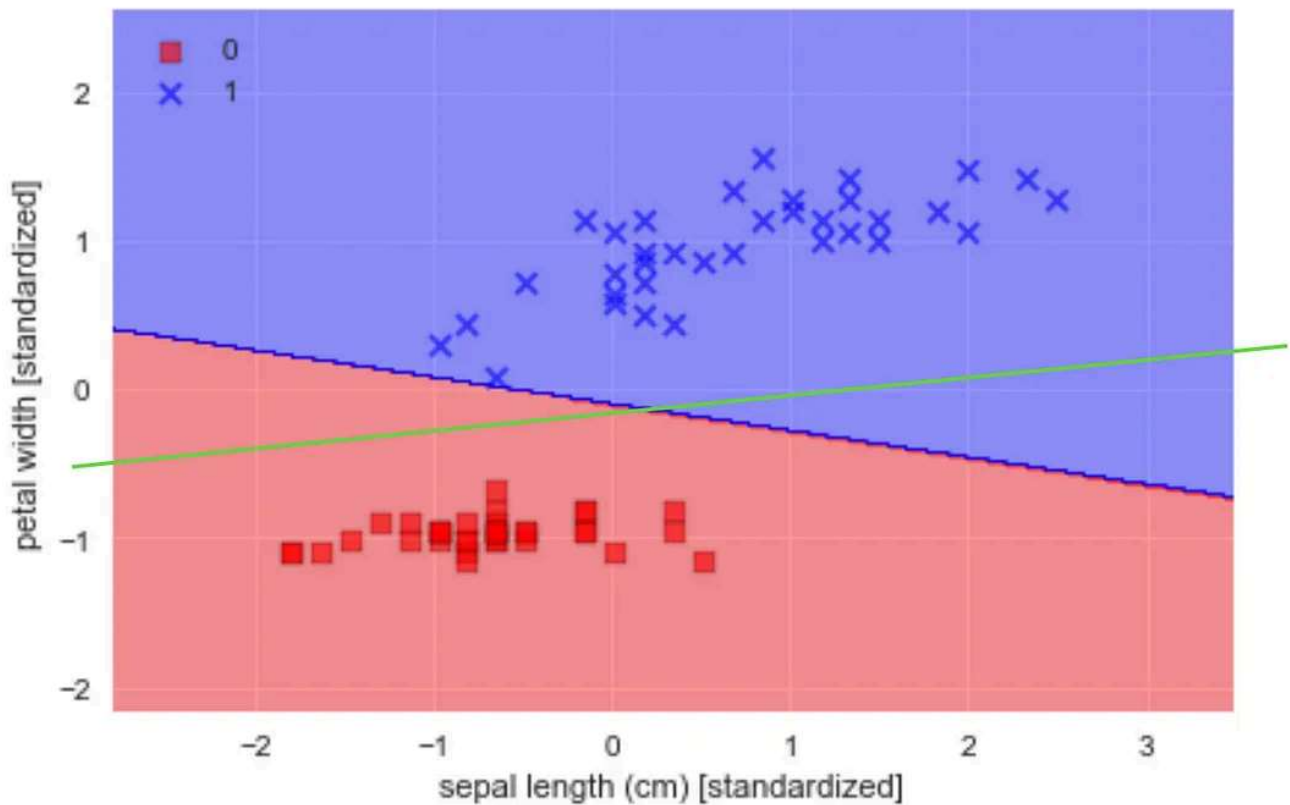
Logistic Regression優點：

1. 資料不需要線性可分
2. 可以獲得A類跟B類的機率

3. 實務上Logistic Regression執行速度非常快

Logistic Regression缺點：

1. 線的切法不夠漂亮，以人的觀察應該要大概要像是綠色的線才是一個比較好的分法 (下一章的SVM將會解決這個問題)



程式碼

感謝你閱讀完這篇文章，如果你覺得這些文章對你有幫助請在底下幫我拍個手（長按最多可以拍50下手）。

[Python資料分析&機器學習]這系列文章是我在Hahow上面所開設課程的講義，如果你是新手想著看影片一步一步學習，可以參考這門課：<https://hahow.in/cr/pydataml>

如果你對什麼主題的文章有興趣的話，歡迎透過這個連結告訴我：

<https://yehjames.typeform.com/to/XIIVQC>

有任何問題也歡迎在底下留言或是來信告訴我：yehjames23@gmail.com

參考閱讀

1. [\[書\]Python 機器學習](#)
2. [林軒田 機器學習基石](#)

Python

Machine Learning

Logistic Regression

Data Analysis



Follow

Published in JamesLearningNote

1.7K Followers · Last published Dec 25, 2017

James學習筆記



Follow

Written by Yeh James

7K Followers · 402 Following

Responses (2)



What are your thoughts?

Respond



Owenwang

May 30, 2020



等你來預測芝加哥的天氣

[Reply](#)

YuChou Chen

May 20, 2018

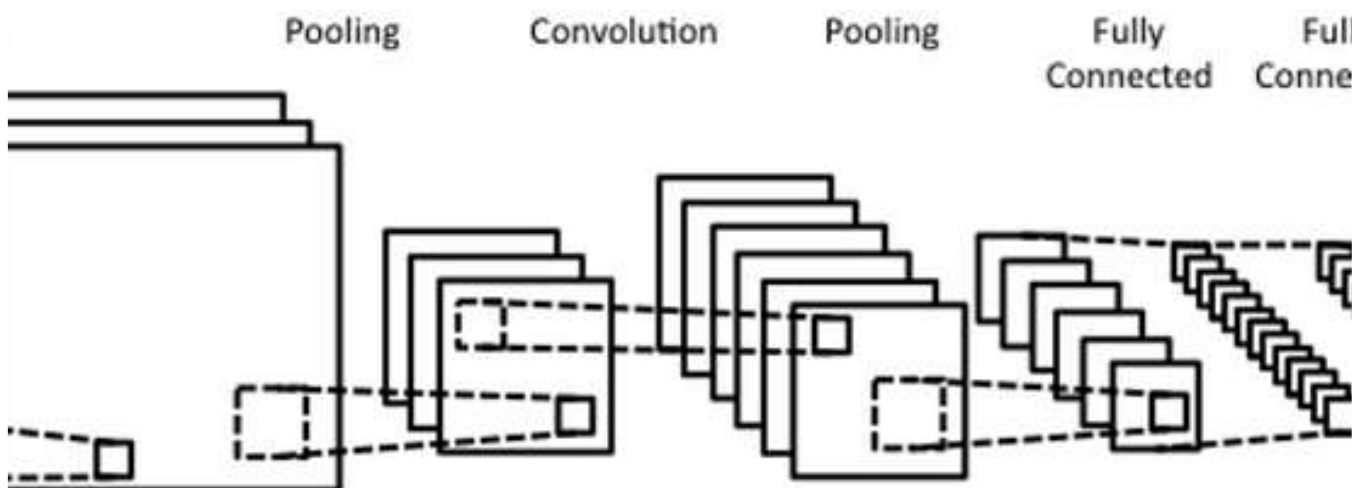


The sigmoid function is a special case of the Logistic function.

Reference link:

<https://stats.stackexchange.com/questions/204484/what-are-the-differences-between-logistic-function-and-sigmoid-function>[Reply](#)

More from Yeh James and JamesLearningNote



In JamesLearningNote by Yeh James

[資料分析&機器學習] 第5.1講: 卷積神經網絡介紹(Convolutional Neural Network)

[資料分析&機器學習] 第3.4講：支援向量機(Support Vector Machine)介紹

Nov 3, 2017

 21K 1



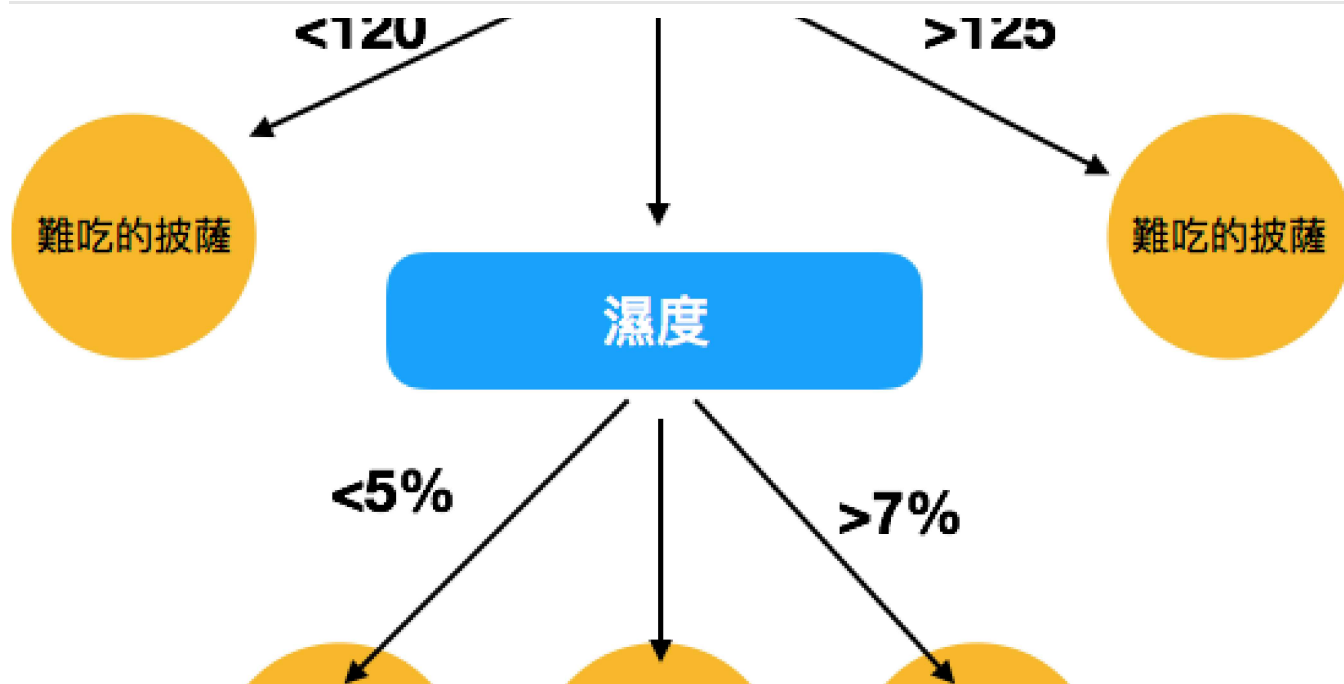


In JamesLearningNote by Yeh James

[資料分析&機器學習] 第3.1講：Python 機器學習以及Scikit-learn介紹

機器學習介紹

Oct 21, 2017 👍 6.9K 💬 1



In JamesLearningNote by Yeh James

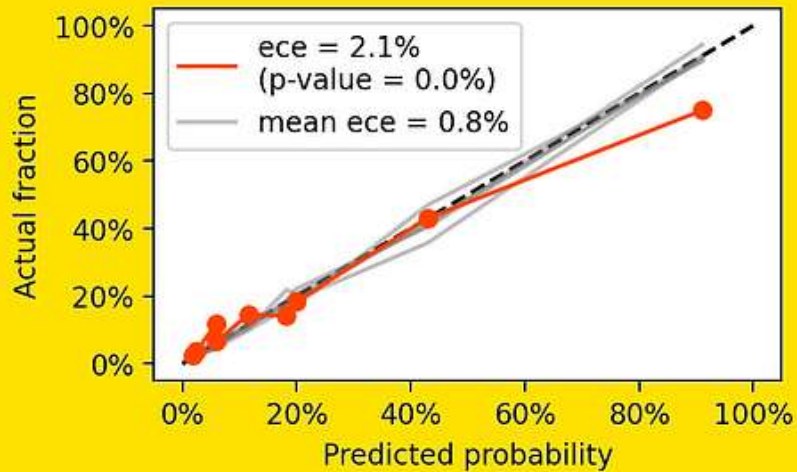
[資料分析&機器學習] 第3.5講：決策樹(Decision Tree)以及隨機森林(Random Forest)介紹

在前面的章節我們說明了如何使用Perceptron, Logistic Regression, SVM在平面中用一條線將資料分為兩類，並且Logistic Regression以及...

Nov 5, 2017 👍 18.7K 💬 3

[See all from Yeh James](#)[See all from JamesLearningNote](#)

Recommended from Medium



In Data Science Collective by Samuele Mazzanti

How to Test if Your Model's Probabilities Are Good (Enough)

Finding a meaningful baseline for the calibration error of any classification model



5d ago



246



3





Jessica Stillman

Jeff Bezos Says the 1-Hour Rule Makes Him Smarter. New Neuroscience Says He's Right

Jeff Bezos's morning routine has long included the one-hour rule. New neuroscience says yours probably should too.



Oct 30, 2024



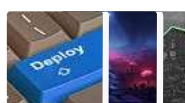
23K



669



Lists



Predictive Modeling w/ Python

20 stories · 1836 saves



Practical Guides to Machine Learning

10 stories · 2212 saves



Coding & Development

11 stories · 1014 saves



Natural Language Processing

1956 stories · 1604 saves

30 Days Data Science Series

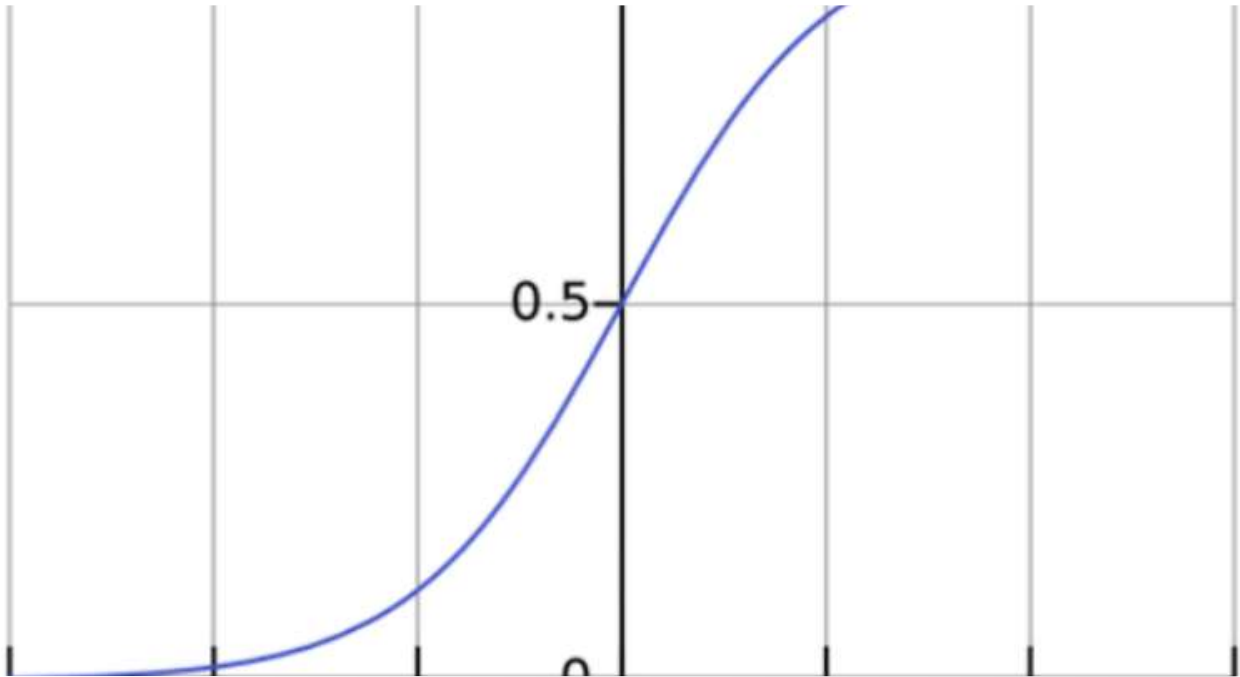


Ime Eti-mfon

Day 02—Logistic Regression

Concept: Binary Classification

Jan 21



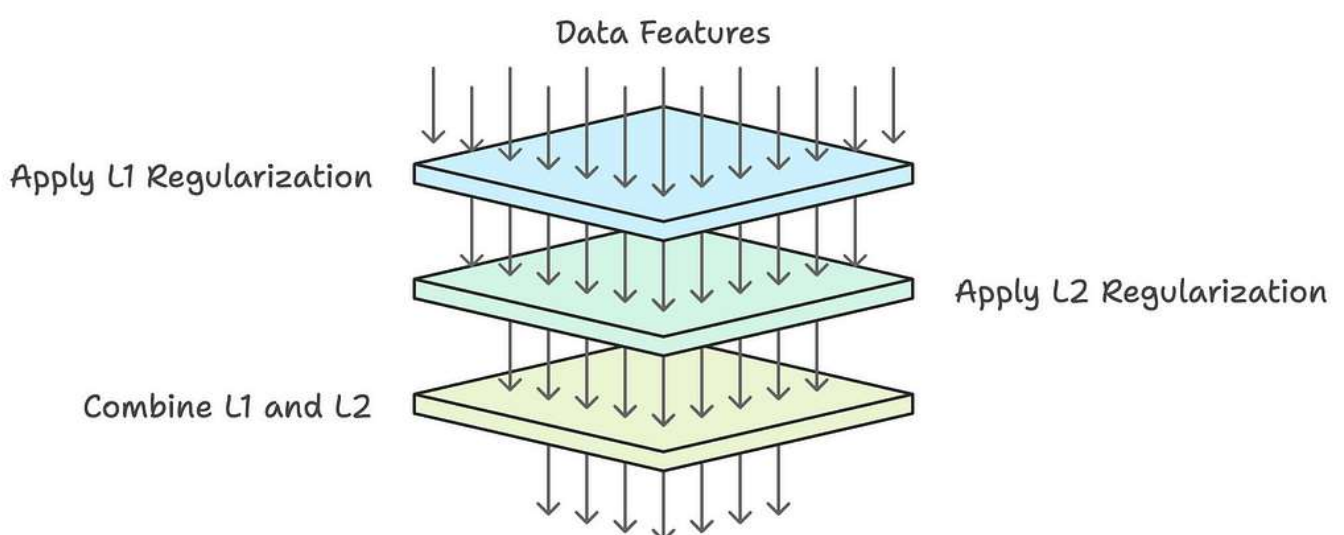
In GoPenAI by Abhay Dodiya

Logistic Regression Python—NIFTY Example

What is Logistic Regression Analysis ? This is often asked question as people are generally familiar with Linear Regression terminology...





Oct 4, 2024



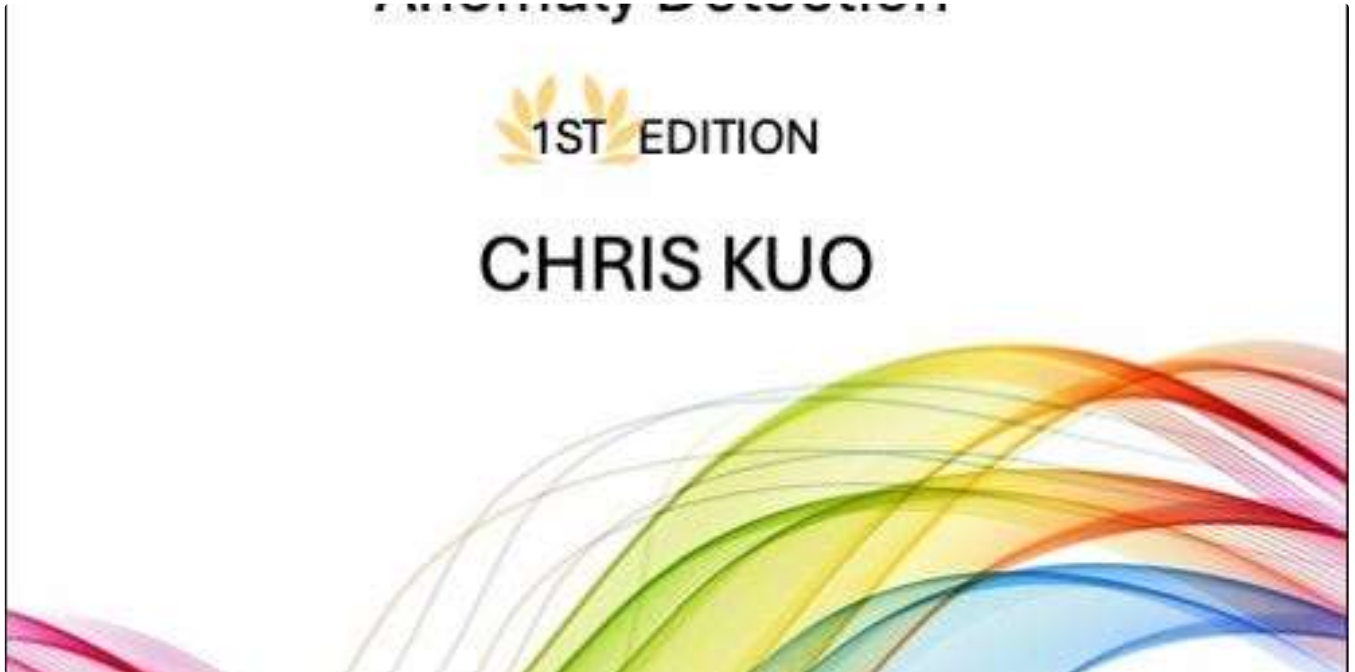


In FunTech Academy by Priyakant Charokar

Chocolate, Toys, and the World of AI: Regularization and Assumptions in Linear Regression

Imagine you're running a toy and chocolate store   , and you want to predict how many chocolates or toys you'll sell based on their...

Dec 10, 2024



In Dataman in AI by Chris Kuo/Dr. Dataman

Monte Carlo Simulation for Time Series Probabilistic Forecasting

Its application on stock market prices



Mar 15, 2024



632



6

[See more recommendations](#)