WARMING UP

1.

Hai bài toán cơ bản

Hãy bắt đầu bằng cách đối mặt với hai bài toán ML đơn giản

1. Tính tiền lương cần trả cho một nhân viên để anh ta thỏa mãn với 75%

“

2. Ra quyết định để cho một khách hàng vay tiền hay không dựa vào phân loại vùng an toàn và vùng rủi ro

TODAY'S TASKS

- Phân biệt được bài toán supervised vs unsupervised và classification vs regression.
- Hiểu và sử dụng mô hình linear regression
- Giải bài toán đầu tiên sử dụng mô hình đã học

SUPERVISED

UNSUPERVISED

Có 3 bài toán cơ bản nhất trong Machine learning: supervised, unsupervised và re-enforcement learning

Supervised learning (có nhãn)

Classification
Phân loại dữ liệu vào các nhóm khác nhau. Ví dụ:
-2 classes: email spam hay không
-10 classes: một số viết tay là số ? (0-9)

Regression
Ước lượng giá trị của một đại lượng.
Ví dụ:
-Dự đoán điểm thi nếu biết thời gian lên lớp, số điểm thi các bài thi trước
-Dự đoán giá trị một cổ phiếu

Unsupervised learning

Clustering
Phân dữ liệu vào các nhóm có chung đặc tính.
Ví dụ: Nhóm khách hàng lại thành các nhóm khác nhau.
Phân biệt với classification.

Dimension reduction
Giảm độ phức tạp (số chiều) của dữ liệu bằng cách bỏ bớt những chiều phụ thuộc.
Ví dụ: Chuyển dữ liệu từ không gian 100 chiều về 10 chiều.

Key feature explanation
Giải thích đặc trưng của dữ liệu.
Ví dụ: Đặc tính của những người thích có chung sở thích

MINI TASK 1

Bạn có thể nhận ra?

Hãy xem xét các bài toán thực tế sau

Khách hàng nào là khách hàng hài lòng?

Source: Kaggle.com

Prize: 60,000$

- Mức độ thỏa mãn của khách hàng là phương thức đánh giá quan trọng của sự thành công
- Những khách hàng không thỏa mãn sẽ rời bỏ bạn mà hiếm khi để lại lời phàn nàn
- Hãy tìm ra nhóm khách hàng có dấu hiệu không thỏa mãn (nhóm khách hàng sẽ rời bỏ bạn)

Đâu là CPC fraud click?

- Click có thể khiến một công ty lãng phí một khoản tài chính lớn cho quảng cáo.
- Một hệ hống quảng cáo có thể có tới 30% ~ 50% là click “ảo”
- Làm sao để “detect” được đâu là click ảo (fraud click)

Nén một ảnh xuống kích thước nhỏ hơn

- Những bức ảnh được upload lên facebook đều được nén lại để giảm dung lượng lưu trữ
- Làm sao để giảm kích thước ảnh mà vẫn giữ được chất lượng ảnh ở kích thước mới

High quality JPEG
File Size: 77.9 kb

Medium quality JPEG
File Size: 19.11 kb

LINEAR REGRESSION

Ước lượng giá trị của một đại lượng với mô hình tuyến tính

What is linear regression

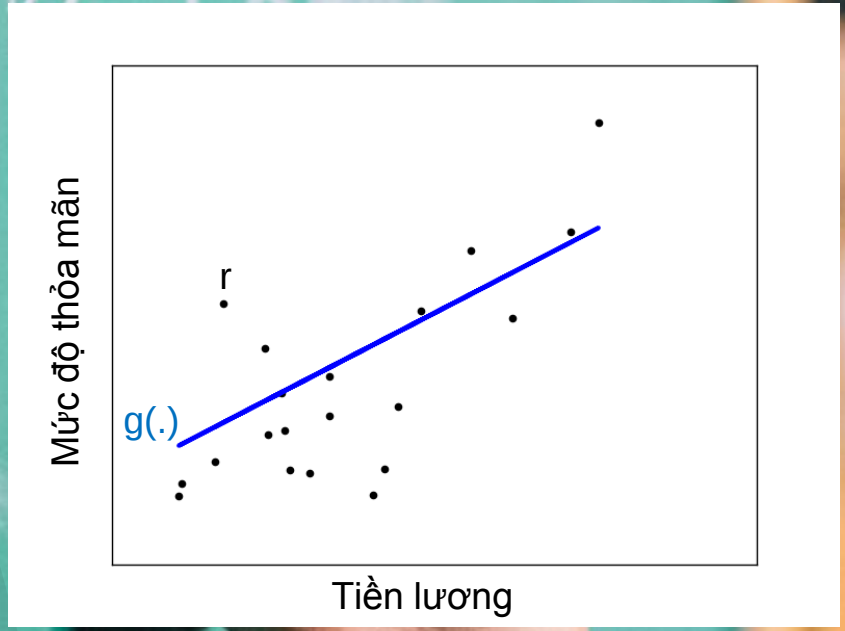
$$r = f(x) + \varepsilon$$

$$f(x) \approx g(x|\theta)$$

$$E(\theta|\mathrm{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g(x^t|\theta)\right]^2$$

$$Linear: g(x|\theta) = g(x|\mathrm{w}_1, \mathrm{w}_0) = \mathrm{w}_1 x^t + w_0$$

r – giá trị của label
f(x) – hàm ước lượng
ε – sai số
g(x|θ) – model
E – sai số



MINI TASK 2

Load tập dữ liệu diabete trong scikit-learn

Sử dụng python, import thư viện scikit-learn để load tập dữ liệu. Tìm số samples? Kích thước không gian feature?

MINI TASK 3

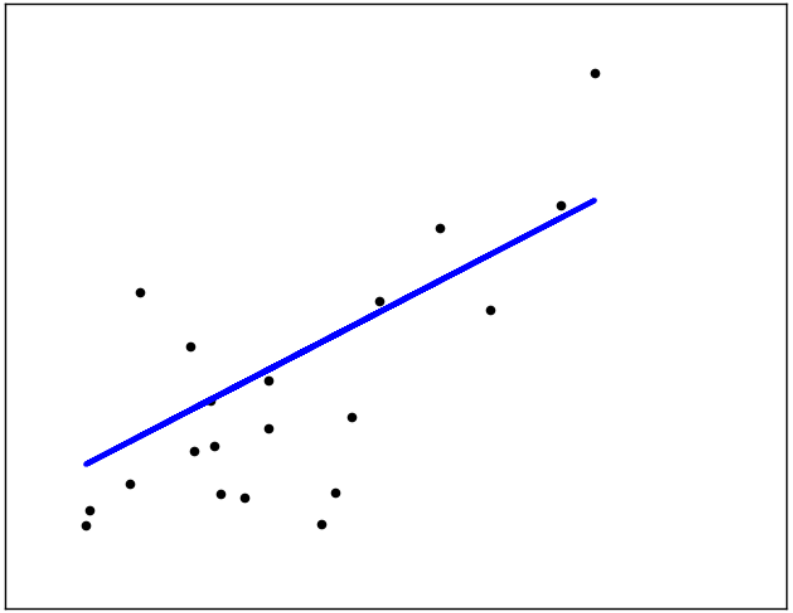
Sử dụng thư viện matplot lib để vẽ đồ thị

Sử dụng thư viện matplot lib để vẽ hai kiểu dữ liệu: scatter và đồ thị

DIRTY HAND

Bài toán regression

Sử dụng python và scikit-learn để giải quyết bài toán regression với mô hình linear regression




$alary

Quy trình giải quyết bài toán

MINI TASK 4

Sử dụng mô hình linear regression

Hãy sử dụng mô hình linear regression để tính được mức độ thỏa mãn của một nhân viên tùy theo mức lương mà nhân viên đó được nhận

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn import datasets, linear_model

# Load the diabetes dataset
diabetes = datasets.load_diabetes()


# Use only one feature
diabetes_X = diabetes.data[:, np.newaxis, 2]

# Split the data into training/testing sets
diabetes_X_train = diabetes_X[:-20]
diabetes_X_test = diabetes_X[-20:]

# Split the targets into training/testing sets
diabetes_y_train = diabetes.target[:-20]
diabetes_y_test = diabetes.target[-20:]

# Create linear regression object
regr = linear_model.LinearRegression()

# Train the model using the training sets
regr.fit(diabetes_X_train, diabetes_y_train)
```

Source: http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html
Code: http://scikit-learn.org/stable/_downloads/plot_ols.py

```
# The coefficients
print('Coefficients: \n', regr.coef_)
# The mean square error
print("Residual sum of squares: %.2f"
      % np.mean((regr.predict(diabetes_X_test) - diabetes_y_test) ** 2))
# Explained variance score: 1 is perfect prediction
print('Variance score: %.2f' % regr.score(diabetes_X_test, diabetes_y_test))

# Plot outputs
plt.scatter(diabetes_X_test, diabetes_y_test,  color='black')
plt.plot(diabetes_X_test, regr.predict(diabetes_X_test), color='blue',
         linewidth=3)

plt.xticks(())
plt.yticks(())

plt.show()
```

Source: http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html
Code: http://scikit-learn.org/stable/_downloads/plot_ols.py

Statistics

185,244 samples

And a lot of samples

90%

Total success!

OUR PROCESS IS EASY

Xử lý dữ liệu

Chọn lựa model

Training và testing

LET'S REVIEW SOME CONCEPTS

Supervised and unsupervised
Bài toán có nhãn và không có nhãn

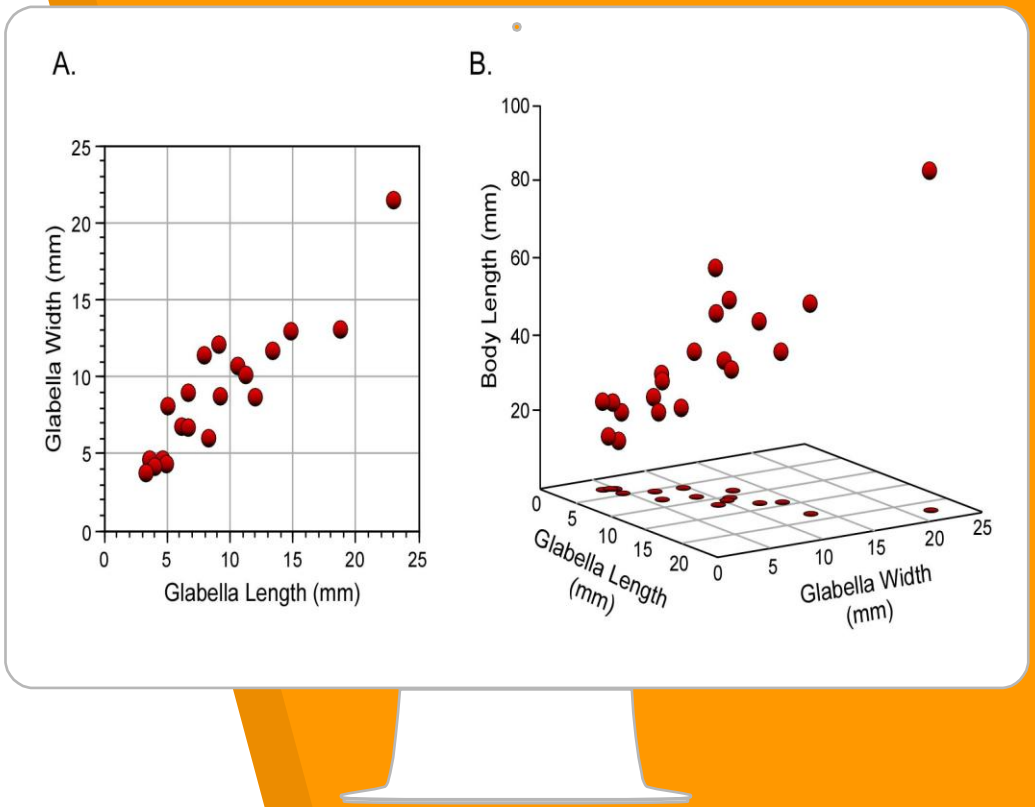
Classification and regression
Bài toán phân lớp và bài toán tìm giá trị (kỳ vọng)

Linear regression
Mô hình regression tuyến tính

Clustering and dimension reduction
Bài toán phân cụm và giảm số chiều dữ liệu

EXTRACT TASK

Load dữ liệu từ file, Sử dụng mô hình linear regression cho bộ features nhiều chiều



THANKS!

Any questions?

You can find me at
caothanhha9@yahoo.com/gmail.com