Homework 4

`TA mail: itai.mmcvlab@gmail.com`

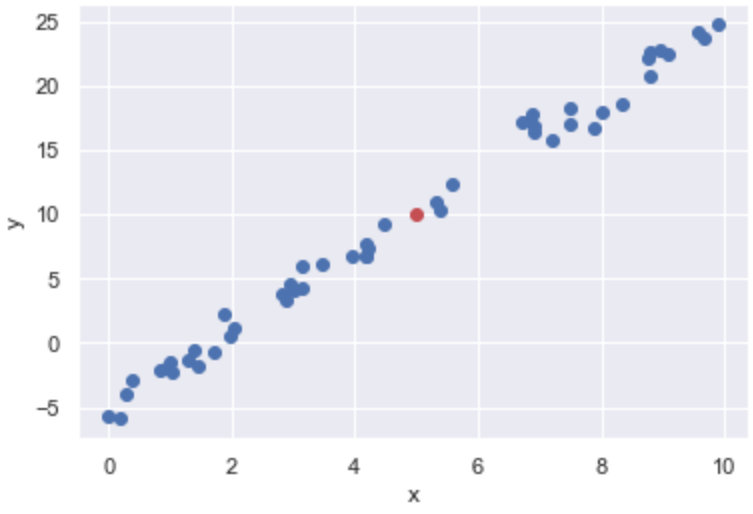
Linear Regression Problem

- 線性迴歸是統計上在找多個自變數和依變數之間的關係所建出來的模型
- 只有一個自變數(x)和一個應變數(y)的情形稱為簡單線性(simple linear regression)迴歸
- 大於一個自變數(x1,x2,...)的情形稱為多元迴歸(multiple regression)

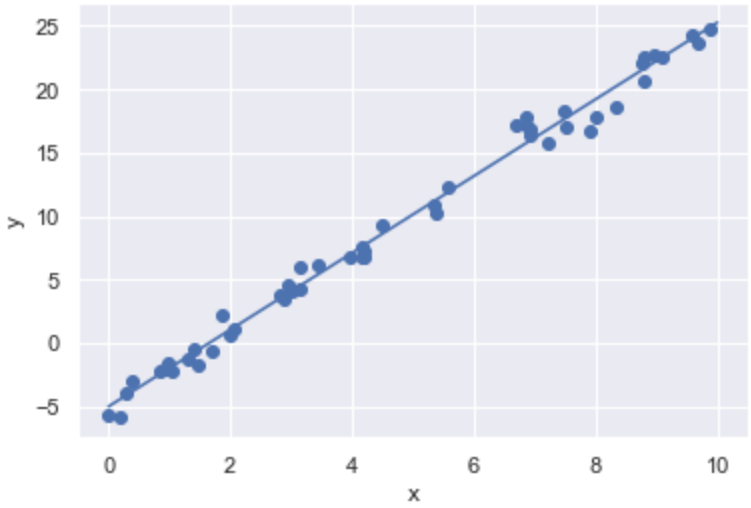
reference: https://ithelp.ithome.com.tw/articles/10243284

Example of simple linear regression

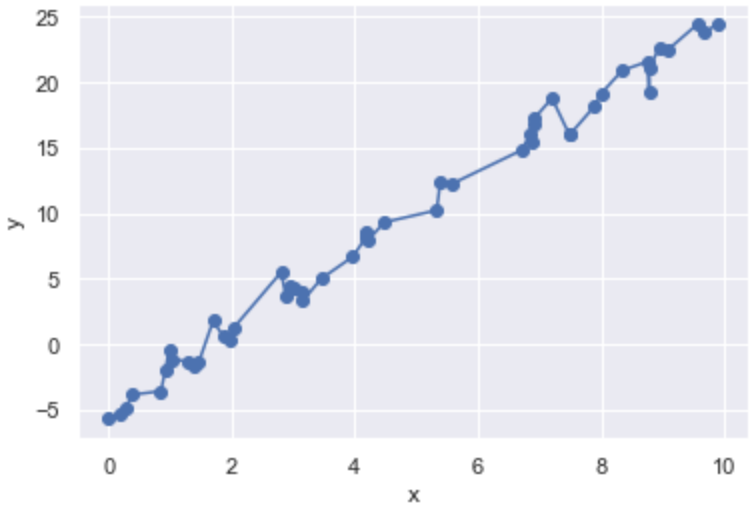
- 下圖為一些資料點，每個點都有其所對應的x, y座標，以下圖紅色點為例，其x, y座標為(5, 10)。



- 目標：盡可能找出一條線來擬合下圖的資料點。



- 當然你也可以直接把所有點連接起來，但這可能會有 overfitting 的問題，所以如何適當的找出線是一件重要的事情。



1. reference: https://ithelp.ithome.com.tw/articles/10206114

Homework

Descriptions

- 本次作業主要是要實作出 Nonparametric regression，詳細內容可以參考老師投影片 chapter 19. slide no. 88~92。
- 這次作業同學需要至少完成老師上課有提到的兩個 nonparametric regression 的方法：　　（如果同學有想實作其他更複雜的方法也可以）
  - k-nearest-neighbors linear regression
  - Locally weighted regression
- 助教會提供兩個 dataset，分別為 data1.npz 和 data2.npz（以下會對這兩個多做講解），同學要使用 nonparametric regression 的方法去擬合這兩個 dataset（其中一個會是比較簡單的 dataset，另一個是比較難一點的 dataset）
- 同學們拿到的只會是兩個 training dataset，並不是整個 dataset，所以要小心 overfitting。

Dataset

- data1.npz - simple regression dataset
- 每個點的x, y 座標 （x 為自變數、y為應變數）

```
# example usage of data1.npz
data = np.load('dataset/data1.npz')

# <class 'numpy.ndarray'>
# shape: (1000,)
# x
data['X'] = [1, 9, 2, 5 ... 2]

# <class 'numpy.ndarray'>
# shape: (1000,)
# y
data['y'] = [7, 12, 7, 3 ... 1]
```

- data2.npz - multiple regression dataset
- 每個點的 x0, x1, y 座標（x0, x1 為二個獨立的自變數、y為應變數）

```
# example usage of data2.npz
data = np.load('dataset/data2.npz')

# <class 'numpy.ndarray'>
# shape: (1000, 2)
# x0, x1
data['X'] = [[1, 2], [1, 7], [9, 5] ... [4, 10]]

# <class 'numpy.ndarray'>
# shape: (1000,)
# y
data['y'] = [7, 12, 7, 3 ... 1]
```

Require

- Code for k-nearest-neighbors linear regression (25%)
  - 用 python 實作 k-nearest-neighbors linear regression
  - 命名為 knn.py
- Code for Locally weighted regression (25%)
  - 用 python 實作 Locally weighted regression
  - 命名為 lw.py
- Code for other method (option) (5%)
  - 用 python 實作其他 Nonparametric regression 的方法
  - 不規定命名
- Report (40%)
  - 說明各個方法的實作與結果、比較不同參數對於 regression 的影響等等……
  - 命名為report.pdf
- Requirement (5%)
  - 將有使用到的 package 列出來，可以用 `pip freeze > requirements.txt` 自動生成

- 命名為 requirements.txt

```
# example of requirements.txt
matplotlib==3.5.1
numpy==1.18.5
pandas==1.0.4
```

Notice

- 請使用 python 完成作業
- 不可以直接用別人提供的演算法（不直接 call function）
  - 舉例來說，不可以直接 `from sklearn.neighbors import KNeighborsClassifier`
  - 可以使用 numpy 做簡單 array 運算
  - 可以使用 matplotlib 做視覺化
  - 可以使用 python 內建函式庫，詳細請見，https://docs.python.org/zh-tw/3.7/tutorial/stdlib.html
  - 其他助教沒有提到的 package 基本上不能使用，如有任何問題歡迎寄信詢問助教
- 檔名記得取正確
- 禁止抄襲
- 繳交格式：將程式碼以及報告一起進行壓縮，並命名為 `hw4_學號.zip`

```
hw2_P12345678.zip
|- knn.py
|- lw.py
|- other_methods.py
|- reports.pdf
|- requirements.txt
```
