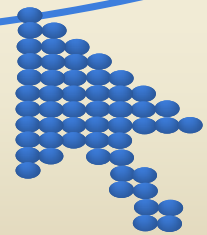
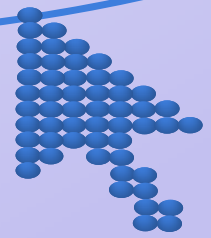
第六、七、八週

3. 資料探勘與機器學習(9hr)



3.1 相關分析與迴歸

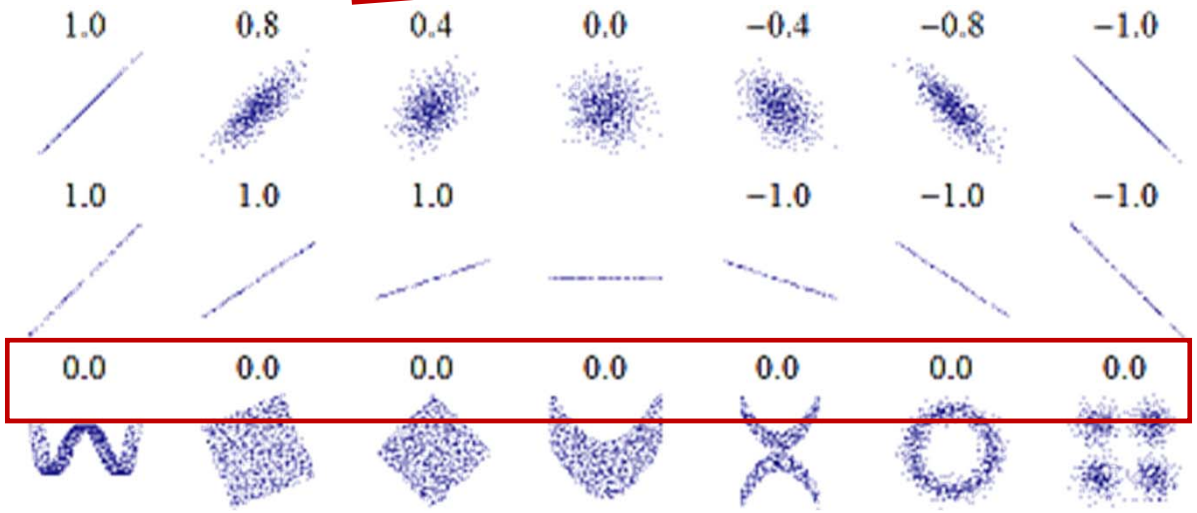


從單變量到多變量

相關分析(續)

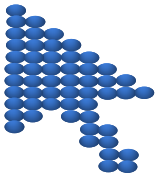
- 通常在研究線性關係時，都會先以散佈圖 (scatter plot) 輔助觀察資料走勢。然而散佈圖僅反映兩個變數之間的相互關係，無法明確定義其關係強度

2D空間兩變量散佈狀況與Pearson相關係數值



真的無關嗎？

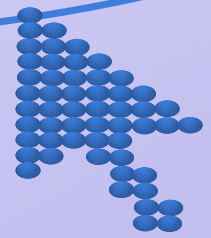
迴歸分析(續)



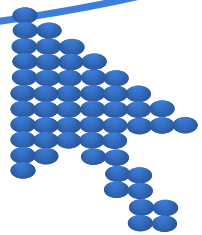
- 羅吉斯迴歸 (Logistic Regression)
  - 當應變數是分類變數，無法使用前述的線性迴歸模式，可使用羅吉斯迴歸來分析：

$$log\left(\frac{P(Y_i=1)}{P(Y_i=0)}\right)=\log\left(\frac{\pi}{1-\pi}\right)=\beta_0+\beta_1X_{i,1}+\cdots+\beta_1X_{i,p-1}+\varepsilon_i=X_i{}^T\beta$$

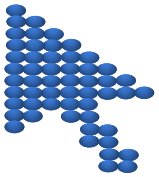
3.3 機器學習應用



支援向量機的應用



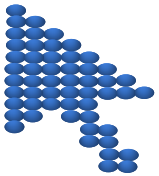
支援向量機分類



- 支援向量機(Support Vector Machines, SVM) 是分類、異常偵測與迴歸的常用工具之一。
- SVM 最早是 Vladimir Vapnik 在 1960 年代中期發展出來的一系列統計模型，其分類模型起源於最大邊界分類器(maximal margin classifiers)，藉由最大化分類超平面與資料之間的邊界幅度，決定出分割不同類樣本的最佳決策邊界。
- 2D空間線性不可分的樣本經過核函數的轉換，在3D空間為線性可區分，即低維度轉換至高維度空間。

參考: https://youtu.be/OdlNM96sHio

- 目標變數 letter 的分佈仍然是吾人應該關心的，結果顯示樣本在 26 個英語字母間散佈相當平均。

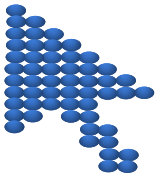


```
# 各整數值變數介於0 到15 之間(4 bits 像素值)
print(letters.describe(include = 'all'))

#         letter           xbox  ...          yedge        yedgex
# count    20000  20000.000000  ...  20000.000000  20000.00000
# unique      26            NaN  ...            NaN          NaN
# top          U            NaN  ...            NaN          NaN
# freq       813            NaN  ...            NaN          NaN
# mean       NaN       4.023550  ...       3.691750      7.80120
# std        NaN       1.913212  ...       2.567073      1.61747
# min        NaN       0.000000  ...       0.000000      0.00000
# 25%        NaN       3.000000  ...       2.000000      7.00000
# 50%        NaN       4.000000  ...       3.000000      8.00000
# 75%        NaN       5.000000  ...       5.000000      9.00000
# max        NaN      15.000000  ...      15.000000     15.00000

# 目標變數各類別分佈平均(預設依各類次數降冪排序)
print(letters['letter'].value_counts())
```

```
U    813
D    805
P    803
T    796
M    792
A    789
X    787
Y    786
N    783
Q    783
F    775
G    773
E    768
B    766
V    764
L    761
R    758
I    755
O    753
W    752
S    748
J    747
K    739
C    736
H    734
Z    734
Name: letter, dtype: int64
```



```
# 真假值矩陣 (類似 R 語言的which函數使用 arr.ind=TRUE)
print(np.argwhere(threTF == True))
# [[0 2]
#  [1 3]
#  [2 0]
#  [3 1]]

# 核對變數名稱，注意相關係數計算時已排除掉第 1 個變數letter
# 預測變數第0~3個為 [xbox,ybox,width,height],4個變數之間似乎高度相關
(> 0.8)，對於有母數的建模方法可能造成不良後果，但對支援向量機來說卻不是個問
題(魯棒性佳)。
print(letters.columns[1:5])
# Index(['xbox', 'ybox', 'width', 'height'], dtype='object')
```

```
# pandas 繪製盒鬚圖 boxplot()，選取 xbox, letter
# 圖形顯示變數 x.box(盒子的水平位置)僅影響字母 A、I、J、L、M 和 W 的辨識
ax1 = letters[['xbox', 'letter']].boxplot(by = 'letter')
fig1 = ax1.get_figure()
# fig1.savefig('img/xbox_boxplot.png')
```