# 07/21 DM 上午場

洪子軒

**Sent:** Thursday, July 21, 2016 11:53 AM
**To:** 洪子軒

【分群】

距離：cosine、歐幾里德距離
分幾群最好？沒有標準答案，但可以透過一些指標（投票法）來衡量

＞k-means（平均值）
Sum of Squared Error (SSE) 誤差平方和（越小越好）
期望：群間差異（大），群內差異（小）
http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html

```
reduced_data = PCA(n_components=2).fit_transform(data)
```

＞k-medoids 中位數（樣本點代表）
找某一成員和其他人距離平方和最小
https://github.com/salspaugh/machine_learning/blob/master/clustering/kmedoids.py

＞Fuzzy c means （FCM）
權重的概念，當樣本點距離群中心越遠，可能性/權重越小
遞迴時，中心點和權重都會加進去影響每次的結果
指標（分幾群比較好，值越小越好）：cmearns、FS、XB、PC（partition coeeficient）、PE（partition entropy）、DB
※FS、XB 較常用

--
洪子軒 Tzu-Hsuan Hung
中華電信研究院 巨量資料所
TEL: (03)-4245128
Email: Lucas@cht.com.tw
32661桃園市楊梅區電研路99號
--