16/6/20

Assignment 08

Yuval S. Katz
204025258

1) a- ex8_sec1.png + ass_1.py.

b- PC1 dimension spans 2 main clusters around [0-30] and [-30,-10]. with 2 centers of dense prob. (0, 0.5)

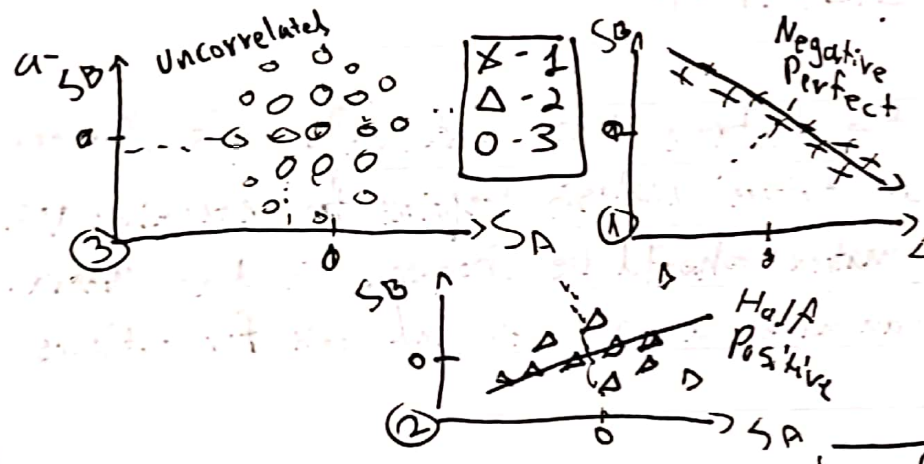
PC2 dimension spans data around [0-15] with no clear division.

c- PC1 explains 58.3%
PC2 explains 37.2% } of the variance, together 95.5% => of the variance in the data can be explained using PC1+2

1-95.5% remains unexplained by this representation

d- Vector 13 has max_var, with entropy of 2.84 (out of 20) bits on avg to pass information
PC1 explains has high variance, with " " 2.83 "
I used 10 bins for the dist. calculations. Since PC1 gets most of the attribute of the data orthogonaly (dimension) to the features, it uses a similar dist. to represent the data, as do vector/feature 13. thus, the same (or at least very close) entropy, (regardless of bin size, just to represent the data to a sufficient level)

2) a-



$cov_1(S_A, S_B) = -1$
$cov_2(S_A, S_B) = 0.5$
$cov_3(S_A, S_B) = 0$

b- ass_2.py
Simulation of values with 1-3 cov:

①	②	③
redundent! the same parameter	$\lambda_1 = 1.5$ $\lambda_2 = 0.5$ $V_1 = \binom{0.7}{0.7}$ $V_2 = \binom{-0.7}{0.7}$	redundent! ~~$\lambda_1 = 1$ $V_1 \binom{1}{0}$~~

c- The first (1) is redundent because S_B dosen't add information beyond S_A information. (cov=-1). it sufficient to look at either S_A/S_B alone!

The last (cov=0,0) display the highest redundancy because S_A and S_B are unrelated (=uncorrelated) and hence, adds a lot of inform.

3) a- see file ex8_ass3.py + png.

b - on the fig

c - on the fig → red black purple blue green

centroids (-7,-3) (-5,5) (0,-7) (5,-7) (6,3) } After 60 iterations

4) EM GM.

a- see file ass_4.py+.png

b- $\mu_i, \sigma_i \Rightarrow$ μ_1, σ_1 , μ_2, σ_2 $\mu_3\ \sigma_3$ $\mu_4\ \sigma_4$

$k=i$ (4.35, 2.8) (1.05, 0.87) (4.02, 5.07) (3.9, 4.83)

c- on fig

d- on fig ass_4_3.png

e) That depends on the distance matrix. Visually, k=3 clusters fit the data well, but future analysis requires to determine vs. k=4, and a distance matrix should be chosen. (Eclidian / Manhatan. ....) for example, we can make a DBSCAN and look for outliers ...