ECE/CS 498 DSU/DSG Spring 2020
In-Class Activity 3

NetID: _____

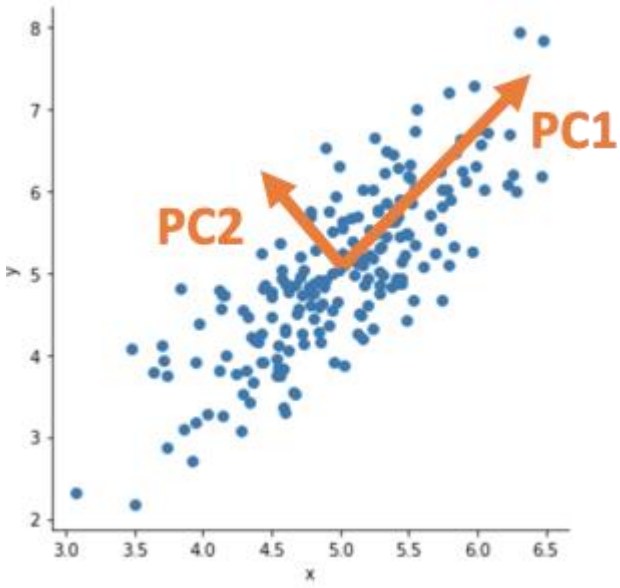
The purpose of the in-class activity is for you to:

(i) Review concepts of principal component analysis
(ii) Go through basic K-means cluster criterion, applications, and caveats

Principal Component Analysis

Problem 1

In the figure below, a scatter plot is provided for data drawn from a multivariate Gaussian distribution. Sketch and label PC1 (principal component 1) and PC2 (principal component 2) on the plot. The arrows should originate from the mean of the distribution and the length of the arrow should specify the variance of the corresponding PC (a rough estimate is fine).



Problem 2

Given the covariance matrix

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

a. Complete the covariance matrix.

b. Compute the eigenvalues of Σ.
Compute eigenvalues by solving $|\Sigma - I\lambda| = 0$

5.83, 2, 0.17

c. Find the first and the second principal components.
Find the principal components by solving for the vectors that satisfy $\Sigma\vec{x} = \lambda\vec{x}$

(0.38, -0.92, 0),
(0, 0, 1),

d. The third principal component is [0.92, 0.38, 0]. Are the principal components orthogonal to each other?
Yes. We can take the pairwise dot product of any of the principal components to see that they are in fact orthogonal to each other.

e. What fraction of total variance does the first principal component account for?
Trace(Σ) = 5.83 + 2 + 0.17 = 8

5.83/8 = 72.86%

Problem 3

The Boston housing dataset consists of 506 samples of Boston housing with their price. Though there are 13 features measured from each sample, for the purpose of this ICA we will only consider the first 3. They are 1) per capita crime rate by town (CRIM), 2) proportion of residential land zoned for logs over 25,000 sq.ft (ZN), and 3) proportion of non-retail business acres per town (INDUS). We have calculated the covariance matrix and part of the correlation matrix for you:

covariance matrix

73.9865782	-40.21595603	23.99233881
-40.21595603	543.93681368	-85.41264806
23.99233881	-85.41264806	47.06444247

correlation matrix

1.0000000	-0.20046922	0.40658341
-0.20046922	1	-0.53382819
0.40658341	-0.53382819	1.000000

Use an online calculator to calculate eigenvalues and eigenvector: https://www.arndt-bruenner.de/mathe/scripts/engl_eigenwert2.htm

a. Complete the correlation matrix (using the formula of correlation and the property that correlation matrix is symmetric. Do not use an online calculator for this question).

b. Find the eigenvalues of the covariance matrix. What is the percentage of total variance explained by the first principal component?

Real Eigenvalues: { 26.223689387745797 ; 76.65240162841977 ; 562.1117433338345 }

562.1117433338345/ (26.223689387745797 + 76.65240162841977 + 562.1117433338345) = 0.84

c. Find the three principal components of the data using the covariance matrix.

for Eigenvalue 26.223689387745797:
[-0.3595825291639235 ; 0.12463386380464005 ; 0.9247522937052941]

for Eigenvalue 76.65240162841977:
[0.9288489216615824 ; 0.14244795237083308 ; 0.34197698985965785]

for Eigenvalue 562.1117433338345:
[0.08910715711006229 ; -0.9819241217418377 ; 0.16698782498518605]

d. Find the eigenvalues of the correlation matrix. What is the percentage of total variance explained by the first principal component?

Real Eigenvalues: { 0.4155573958430107 ; 0.8081855123604621 ; 1.7762570917965275 }

1.7762570917965275 / (0.4155573958430107 + 0.8081855123604621 + 1.7762570917965275) = 0.60

e. Find the principal components of the data using the correlation matrix.

for Eigenvalue 0.4155573958430107:
[-0.3254908288972064 ; 0.5743926120752808 ; 0.7510851133507744]

for Eigenvalue 0.8081855123604621:
[-0.8080509181105482 ; -0.5814820528913279 ; 0.09451103589521698]

for Eigenvalue 1.7762570917965275:
[0.4910289543851363 ; -0.5761525400087777 ; 0.6534055529277919]

f. Observe the contribution of the different features to the first principal component in the two cases above. What can you conclude?

Performing PCA with the covariance matrix results in principal components that greatly emphasize the features with greatest variance. This makes sense since the covariance matrix is more sensitive to changes in scales between features. Since the correlation matrix is the normalized covariance matrix, the correlation matrix emphasizes the different features more equally.

Clustering

Problem 1

As an animal lover, John records the location each time a squirrel appears. These observations are recorded in the 3-D coordinates. Assume you want to cluster 8 observations into 3 clusters using K-Means clustering algorithm with the Euclidean distance measure. After the first iteration of clustering, C1, C2, C3 have the following observations:

C1	*C2*	*C3*
(5,4,6) (4,4,4) (6,6,7)	(7,2,2) (10,1,0)	(5,1,12) (3,2,11) (3,2,10)

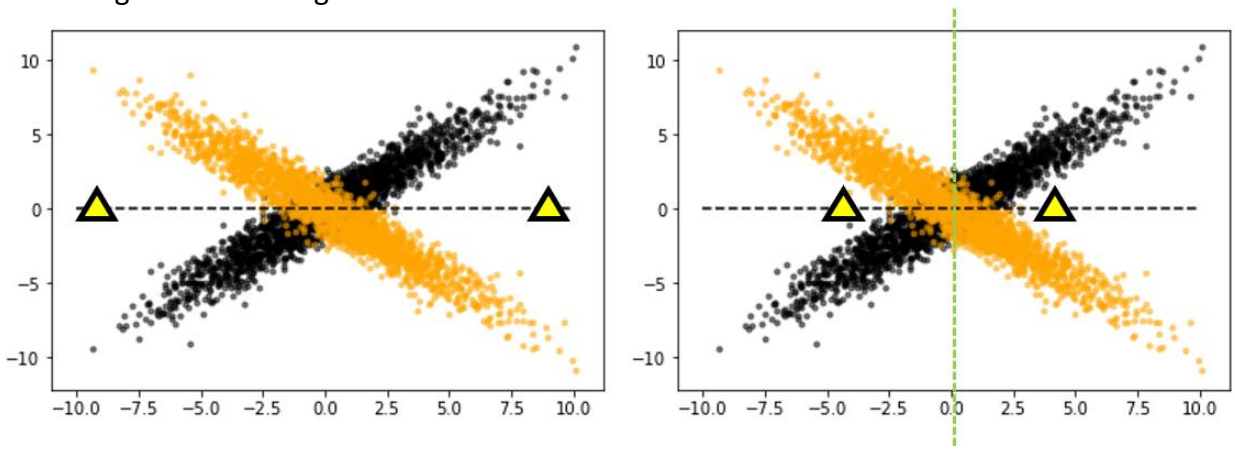
a. What will be the cluster centroids after the second iteration?
C1: ((5+4+6)/3=5, (4+4+6)/3=14/3, (6+4+7)/3=17/3)
C2: ((7+10)/2=17/2, (2+1)/2=3/2, (2+0)/2=1)
C3: ((5+3+3)/3=11/3, (1+2+2)/3=5/3, (12+11+10)/3=11)

b. Does the algorithm converge after the second iteration? State your reason.
Yes, because the calculated centroids (or the assignments of data) remain the same after the second iteration.

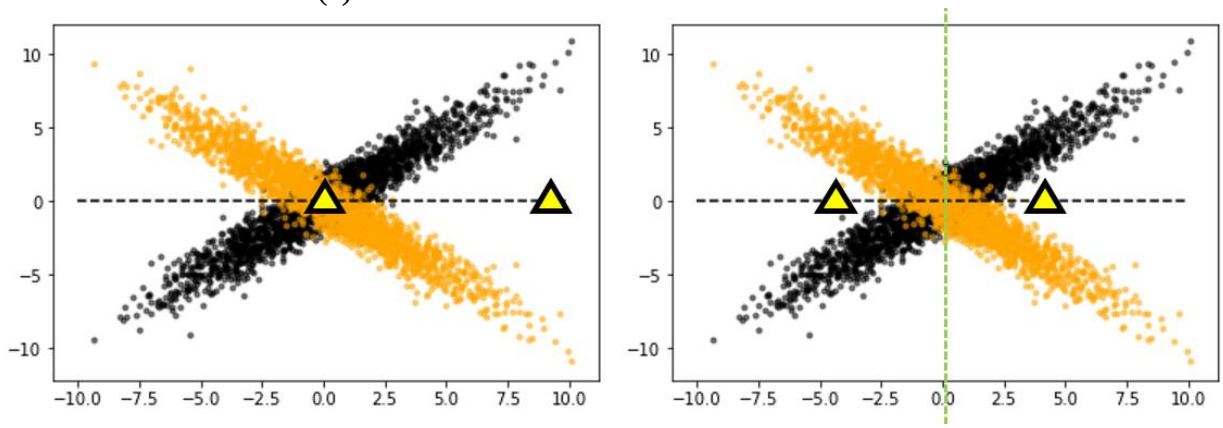
c. Suppose a new observation is at (2,3,5), which cluster (C1, C2, C3) will you assign it to based on the centroids calculated in the question a?
Distance to C1 = 3.496
Distance to C2 = 7.778
Distance to C3 = 6.368
Since (2,3,5) is closest to C1, we assign it to C1.

Problem 2

1. Suppose you are going to perform clustering on a 2-D dataset, and you assigned 2 centroids indicated as triangles on the left figure. (a) and (b) are different cases we put the initial centroids on. Draw the new centroids and decision boundary on the right figure after the K-Means algorithm converges.



(a)



(b)

2. Is K-Means a good clustering method for these data? If not, what clustering method are supposed to have a better result on these data. (Note that the ground truth of these two datasets is that they cross each other and overlap at the origin of the coordinate.)

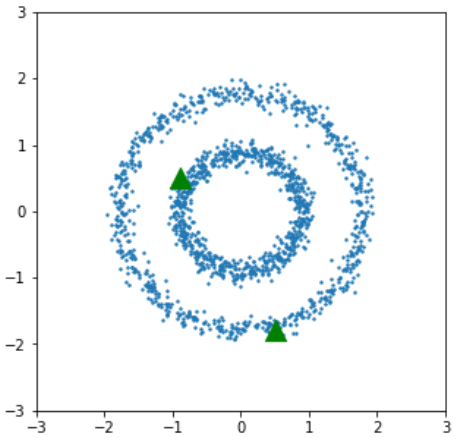
K-Means are not suitable for this dataset since the centroids of two ground-truth clusters overlap each other.

GMM is supposed to have better result.

3. If you still want to do K-Means clustering for these data, how can you transform the data/features so that the two clusters crossing each other can be better identified by the k-means algorithm?

We can take x/y as a new feature, so that the two cluster centers will be at -1 and 1.

4. How can you transform the data in the following figure so that the two ring clusters are identifiable by the k-means algorithm? (Hint: Think about polar coordinates.)



As we are dealing with circles, if we transform our Cartesian (x vs y) coordinates to polar (arc vs radius) coordinates we end up with two distinct rectangular clusters. They have the same arc range but are completely partitioned by their radius.

Problem 3

Assume we have already chosen the parameters of a mixture model with 3 components c_1 ~N(0,1), c_2 ~N(6,2) and c_3~N(3,1). We would like to infer that given a data point x, which component c it is most likely to belong to. To achieve this purpose, we want to infer the posterior distribution $p(c|x)$.

a. Fill in the blank below using the theorem of total probability:

$$p(x) = p(x|c_1)p(c_1) + p(x|c_2)p(c_2) + p(x|c_3)p(c_3)$$

b. Suppose we observe $x = 2$. Given that:

$$\pi_1 = P(c = 1) = 0.4$$
$$\pi_2 = P(c = 2) = 0.3$$
$$P(x|c = 1) = Gaussian(x = 2; \mu_1 = 0, \sigma_1 = 1) \approx 0.054$$
$$P(x|c = 2) = Gaussian(x = 2; \mu_1 = 6, \sigma_1 = 2) \approx 0.027$$
$$P(x|c = 3) = Gaussian(x = 2; \mu_1 = 3, \sigma_1 = 1) \approx 0.242$$

Which component should we assign the observation $x = 2$ to? Justify your choice.

0.3*0.242>0.4*0.054 > 0.3*0.027, so we assign the observation to c_3.