ECE/CS 498 DSU/DSG Spring 2019
In-Class Activity 3

NetID: _____

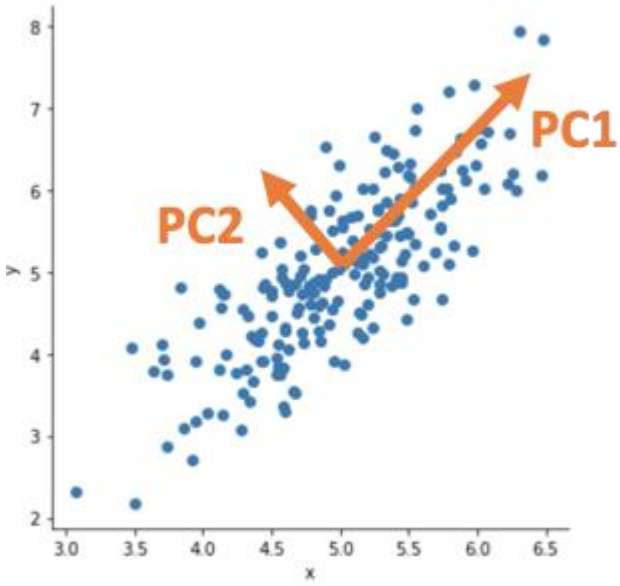
The purpose of the in-class activity is for you to:

(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.

5.83, 2, 0.17

c. Find the first and the second principal components.
(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

e. What fraction of total variance does the first principal component account for?
72.86%

Problem 3

The Iris dataset consists of 150 samples of Iris flower. Four features were measured from each sample: the length and width of the sepals (in centimeters) and petals (in millimeters). The summary statistics are provided as follows:

	sepal_len	**sepal_wid**	**petal_len**	**petal_wid**
mean	5.843333	3.054	37.586667	11.986667
std	0.828066	0.433594	17.644204	7.631607

covariance matrix

0.681122	-0.039007	12.651911	5.134578
-0.039007	0.186751	-3.195680	-1.171947
12.651911	-3.195680	309.242489	128.774489
5.134578	-1.171947	128.774489	57.853156

correlation matrix

1.000000	-0.109369	0.871754	0.817954
-0.109369	1.000000	-0.420516	-0.356544
0.871754	-0.420516	1.000000	0.962757
0.817954	-0.356544	0.962757	1.000000

Use an online calculator: https://www.arndt-bruenner.de/mathe/scripts/engl_eigenwert2.htm

a. Complete the correlation matrix.

b. Find the eigenvalues of the covariance matrix. What is the percentage of total variance explained by the first principal component?
364.042658 3.614271 0.05626 0.250329

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

sepal_len **sepal_wid** **petal_len** **petal_wid**

pc1	0.037547	-0.009341	0.920866	0.387955
pc2	0.051025	-0.045036	0.385190	-0.920324
pc3	0.684110	-0.725493	-0.056267	0.049881
pc4	0.726623	0.686691	-0.021657	-0.002382

d. Find the eigenvalues of the correlation matrix. What is the percentage of total variance explained by the first principal component?
2.910818 0.921221 0.147353 0.020608

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

	sepal_len	**sepal_wid**	**petal_len**	**petal_wid**
pc1	0.522372	-0.263355	0.581254	0.565611
pc2	-0.372318	-0.925556	-0.021095	-0.065416
pc3	-0.721017	0.242033	0.140892	0.633801
pc4	0.261996	-0.124135	-0.801154	0.523546

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

Clustering

Problem 1

Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm with Euclidian distance measure. After the first iteration of clustering, C1, C2, C3 have the following observations:
C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}

a. What will be the cluster centroids if you want to proceed for a second iteration?
C1: ((2+4+6)/3=4,(2+4+6)/3=4)
C2: ((0+4)/2=2,(4+0)/2=2)
C3: ((5+9)/2=7,(5+9)/2=7)

b. Which cluster (C1, C2, C3) will you assign (9,9) to in the second iteration?
C1: sqrt((9-4)^2 + (9-4)^2) = sqrt(50)
C2: sqrt((9-2)^2+(9-2)^2) = sqrt(98)
C3: sqrt((9-7)^2+(9-7)^2) = sqrt(8)
Since (9,9) is closest to C3, we assign it to C3.

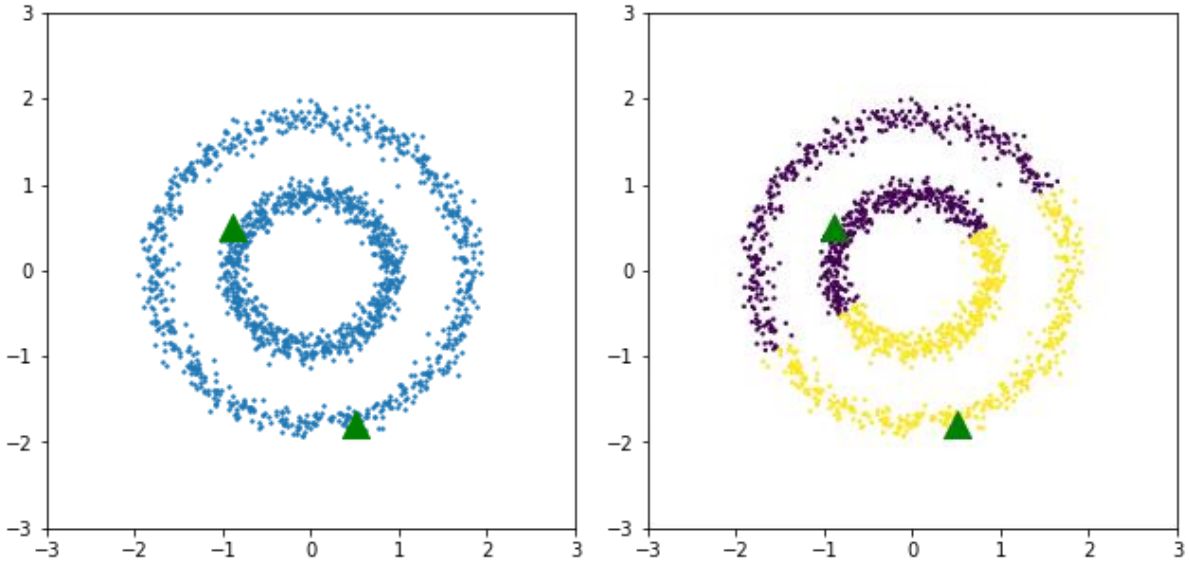
Problem 2

State termination conditions for the k-means clustering algorithm in terms of:

a. Cluster assignment:
   Assignment of observations to clusters does not change between subsequent iterations
b. Change of centroid:
   Centroids do not change between successive iterations
c. SSE:
   Terminate when SSE (sum of squares of errors which represents the distance between data points and corresponding cluster centroid) falls below a threshold.
d. [*Optional*] Number of iterations:
   a pre-specified number of iterations has been reached

Problem 3

Speculate the outcome of applying k-means *with Euclidean distance* on the following plot. Draw the two clusters you think k-means algorithm will identify on the plot when initialized with the seeds as marked. Why do you choose to use or not use k-means in this case?



How can you transform the data 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 radii.

Problem 4

K-means, hierarchical clustering, and GMM fall under the class of _unsupervised_ (supervised/unsupervised) algorithms. Naïve Bayes fall under the class of ___supervised___ (supervised/unsupervised) algorithms.

a. If the solutions you provide to the two blanks above are the same, briefly explain the commonalities; if the solutions you provide are different, briefly explain the differences.
Different; Supervised learning is the machine learning task of learning a function that maps an input to an output based on training input-output pairs; Unsupervised machine learning is the machine learning task of inferring a function to describe hidden structure from "unlabeled" data.

b. Provide additional example(s) for supervised and unsupervised learning algorithms respectively.
Examples of unsupervised learning:
Clustering, Anomaly Detection
Examples of supervised learning:
Regression, Decision Trees, SVM, Naïve Bayes, k-Nearest Neighbors, Neural Networks

Problem 5

Assume we have already chosen the parameters of a model with 2 components c_1 and c_2. 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 Bayes' Rule and the theorem of total probability:

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

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

$$\pi_1 = P(c = 1) = 0.7$$
$$\pi_2 = P(c = 2) = 0.3$$
$$P(x|c = 1) = Gaussian(x = 2; \mu_1 = 0, \sigma_1 = 1) \approx 0.054$$
$$P(x|c = 2) = Gaussian(x = 2; \mu_1 = 6, \sigma_1 = 2) \approx 0.027$$

Which component should we assign the observation $x = 2$ to? Justify your choice.
0.7*0.054 > 0.3*0.027, so we assign the observation to c_2.