Due date: Mar. 10, 2019, 11:59 PM (Arlington time). You have **two** late days — use it at as you wish. Once you run out of this quota, the penalty for late submission will be applied. You can either use your late days quota (or let the penalty be applied). Clearly indicate in your submission if you seek to use the quota.

What to turn in:

1. Your submission should include your complete code base in an archive file (`zip, tar.gz`) and `q1/, q2/,` and so on), and a very very clear README describing how to run it.
2. **A brief report (typed up, submit as a PDF file, NO handwritten scanned copies) describing what you solved, implemented and known failure cases.**
3. Submit your entire code and report to Blackboard.

Notes from instructor:

- Start early!
- You may ask the TA or instructor for suggestions, and discuss the problem with others (minimally). But **all parts of the submitted code must be your own**.
- Use Matlab or Python for your implementation.
- Make sure that the TA can easily run the code by plugging in our test data.

Problem 1

(k-means, **40pts**) Generate 2 sets of 2-D Gaussian random data, each set containing 500 samples using parameters below.

$$\mu_1 = [1, 0],\ \mu_2 = [0, 1.5],\ \Sigma_1 = \begin{bmatrix} 0.9 & 0.4 \\ 0.4 & 0.9 \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} 0.9 & 0.4 \\ 0.4 & 0.9 \end{bmatrix} \tag{1}$$

1. **(20pts)** Write a function `cluster = mykmeans(X, k, c)` that clusters data $X \in \mathbb{R}^{n \times p}$ (n number of objects and p number of attributes) into k clusters. The c here is the initial centers, although this is usually not necessary, we will need it to test your function. Terminate the iteration when the ℓ_2-norm between a previous center and an updated center is ≤ 0.001 or the number of iteration reaches 10000.
2. **(10pts)** Apply your code to the data generated above with $k = 2$ and initial centers $c_1 = (10, 10)$ and $c_2 = (-10, -10)$. In your report, report the centers found for each cluster. How many iterations did it take? Show a scatter plot of the data and the centers of clusters found.
3. **(10pts)** Apply your code to the data generated above with $k = 4$ and initial centers $c_1 = (10, 10)$ and $c_2 = (-10, -10)$, $c_3 = (10, -10)$ and $c_4 = (-10, 10)$. In your report, report the centers found for each cluster. How many iterations did it take? Show a scatter plot of the data and the centers of clusters found.

Problem 2

(Non-parameteric density estimation **60pts**)

1. **(30pts)** Write a function `[p, x] = mykde(X,h)` that performs kernel density estimation on X with bandwidth h. It should return the estimated density $p(x)$ and its domain x where you estimated the $p(x)$ for X in 1-D and 2-D.

2. **(10pts)** Generate $N = 1000$ Gaussian random data with $\mu_1 = 5$ and $\sigma_1 = 1$. Test your function `mykde` on this data with $h = \{.1, 1, 5, 10\}$. In your report, report the histogram of X along with the figures of estimated densities.

3. **(10pts)** Generate $N = 1000$ Gaussian random data with $\mu_1 = 5$ and $\sigma_1 = 1$ and another Gaussian random data with $\mu_2 = 0$ and $\sigma_2 = 0.2$. Test your function `mykde` on this data with $h = \{.1, 1, 5, 10\}$. In your report, report the histogram of X along with the figures of estimated densities.

4. **(10pts)** Generate 2 sets of 2-D Gaussian random data with $N_1 = 500$ and $N_2 = 500$ using the following parameters:

$$\mu_1 = [1, 0],\ \mu_2 = [0, 1.5],\ \Sigma_1 = \begin{bmatrix} 0.9 & 0.4 \\ 0.4 & 0.9 \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} 0.9 & 0.4 \\ 0.4 & 0.9 \end{bmatrix}. \tag{2}$$

Test your function `mykde` on this data with $h = \{.1, 1, 5, 10\}$. In your report, report figures of estimated densities.