# Homework # 4: Unsupervised Learning

**Due: Wednesday, April 21, 2021, 11:59 p.m.**
**Total Points: 100**

This assignment is designed to give you practical programming experience with unsupervised learning. **Please carefully read all instructions below and also periodically check Piazza for updates.**

You have been assigned a partner to work with on this homework assignment (see end of document). A spreadsheet with student contact information is on the Syllabus page of Canvas. Use this information to contact your partner and to establish your working relationship. Submit a post to Piazza (to Instructors) if you have any issues connecting or working with your partner.

Each individual within the pairing should submit a solution to the assignment, to avoid miscommunication issues with your partner. Be sure to specify in the file, whose submission should be graded, as only one will be graded for the pairing. The submitted homework must include your names, as well as all resources that were used to solve the problem (e.g. web sites, books, research papers, other people, etc.). Academic honesty is taken seriously; for detailed information see Indiana University Code of Student Rights, Responsibilities, and Conduct. **Please write a post to Piazza if you are considering using Python libraries that have not been discussed in class.**

**Your assignment must be submitted by committing your code to your IU Github repository before the deadline. Please create a folder called 'hw4' (without quotes) and commit all files for this assignment to this folder.** You are strongly encouraged to submit early and often to avoid any last minute issues and to confirm that you are committing your work correctly. The submission should be in the form of a Python 3 Jupyter Notebook file, along with any special instructions for running the program. **Be sure to run the file before committing, so that we can directly see your results**. Programs that fail to run will not be graded and will result in a zero.

**Late submission policy**: We expect all work to be finished and submitted on time. However, we do understand that there might be times when something unexpected comes up which delays you. Hence, as a late submission policy, we will allow late submissions up to 3 days late, each day carrying a 5 point penalty. If you submit 70 hours after the original deadline, the highest possible score for that assignment will be 85 points, out of 100. If the assignment is submitted 1 minute after the deadline, then a 5 point penalty will be assessed. Late assignments submitted at 72:01 hours after the original deadline will not be accepted. Please, make the arrangements to start and finish your assignments early. This means you should push to Github well before the time stated on the homework. This means you must start the homework as soon as you can and continually commit.

## Question 1. [50 POINTS]

Using the MNIST dataset, perform the following (Note that you may need to perform sampling to reduce the amount of data or perform PCA to reduce the dimensionality of each example beforehand):

1. Perform K-means clustering using an appropriate value for $K$, where the clusters are initialized from random data from the dataset. Discuss how you selected the appropriate value for K.

2. For the appropriate value of K that was selected, display the cluster centroids. Discuss the similarities and differences between the centroids, and how they relate to the labels of the data, if at all.
3. Repeat the above, using k-means++ initialization
4. Now assume that you have labeled information. Randomly choose 1 sample from each of the classes to serve as the initial cluster centers. Run the K-means algorithm. Display the final and initial cluster centroids. Compare and contrast the initial cluster centers to the final cluster centers. Do they still represent the labels?
5. How do the cluster centroids compare for the above three approaches?
6. Compute the variance (as shown in class) for each of the three approaches. How do they compare?

## Question 2. [50 POINTS]

Divide the MNIST dataset into a training and a testing set using 60,000 images for training and the remaining for testing.

- Train a Random Forest classifier using the training data and evaluate how long it takes to train.
- Test the classifier with the testing data and generate a confusion matrix and compute the overall accuracy.
- Use PCA to reduce the dataset's dimensionality, using explained variance ratio's of 95%, 90%, and 85%, respectively.
- Train Random Forest classifiers using the dimensionally-reduced data and evaluate how long it takes to train. Discuss how the explained variance ratio influences training data, along with a comparison to the initial training time.
- Evaluate the classifiers using the testing set, generating the confusion matrices and overall accuracy, for each case. Discuss the performance differences.