San José State University
Department of Applied Data Science

# DATA 220
# Mathematical Methods for Data Analysis

Spring 2021
Instructor: Ron Mak

## Assignment #9

**Assigned:** Thursday, April 15
**Due:** Thursday, April 22 at 5:30 pm
**Team assignment**, 100 points max

**Supervised and unsupervised machine learning**
In this assignment, you will perform both supervised and unsupervised machine learning using Python's Scikit-learn (**`sklearn`**) module.

Explore and try different operations, since Python makes it easy to experiment.

**Unsupervised ML**
Choose a suitable unlabeled dataset that has at least several hundred data points. You can choose a labeled dataset and just ignore the labels for this part of the assignment. Clean and transform the data as necessary, and then load it into a Python data structure.

Use what we did in class with the Iris dataset as a guide:

- Visualize the dataset with a grid of graphs that each plots one data attribute against another. (You do not have to involve all the attributes if there are more than four or five.)
- Perform dimensionality reduction down to two dimensions using both the TSNE and the PCA estimators. Create scatter plots and compare the results of the two estimators and decide which is better (i.e., produces well-separated clusters).
- Perform **k-means clustering**. If you started with unlabeled data, you may need to experiment with different values of $k$. Your graphs of the dimensionally reduced data can help you choose $k$.

> Even though this part of the assignment is unsupervised ML which normally works with unlabeled data, it's OK to "cheat" by using labeled data instead and ignoring the labels during k-means clustering. Then you'll be able to measure how well the clusters match the labels.

After you've performed the k-means clustering, you can label each data point with the cluster to which it was assigned. Now you have labeled data. If you started with labeled data, you can compare the actual labels with your assigned labels.

**Supervised ML**
Split your newly labeled data into training and test data and perform k**-nearest neighbors classification**.

Use what we did in class with the Digits dataset as a guide:

- After training the model, feed it the test data to see how well the model predicts.
- Create a confusion matrix and a classification report using the cluster labels you generated earlier.
- If your dataset was originally labeled, you can also compare the classifications with the original labels.

**Other experiments**
Python makes it easy to explore and experiment with different operations! What additional interesting operations can you perform on the data, either while doing supervised or unsupervised machine learning, or both?

**Report**
Incorporate comments and markdowns in your notebook that describe what you are doing and your interpretation of the results.

**What to submit to Canvas**
Submit your Jupyter notebook into Canvas: **Assignment #9.**

**Rubric**
Your program will be graded according to these criteria:

| Criteria | Max points |
|---|---|
| **Chosen dataset**<br>• Good dataset with any necessary cleanup and transformation. | **10** |
| **Unsupervised machine learning**<br>• TSNE dimensionality reduction and graph.<br>• PCA dimensionality reduction and graph.<br>• Perform k-means clustering. | **40**<br>• 10<br>• 10<br>• 20 |
| **Supervised machine learning**<br>• Split the newly labeled dataset into training and testing data.<br>• Perform k-nearest neighbors classification.<br>• Generate a confusion matrix and classification report. | **40**<br>• 10<br>• 20<br>• 10 |
| **Further experiments**<br>• Any other interesting experiments with the data. | **10** |