Applications of Machine Learning in Remote Sensing
Homework 1

John Smith – johnsmith@rit.edu

https://github.com/johnsmith/repo.git

- In your submission, include **explanation**, **results**, and **the code** for the problem in the same PDF file in form of a Jupyter Notebook Results. Also *separately*, attach solution's codes so I can replicate your results.
- Show your understanding of the problem by providing **explanation**.
- Provide sufficient commenting in your code.
- Ensure all text/images are legible and organized.
- Ensure that your code can reproduce the submitted results.

Create a directory in your repository and name it `eda`, if you already do not have. The workflows and the scripts created in this homework would go under `eda`.

Problem 1

We are using the 12 bands of the Sentinel-2 satellite over Rochester captured last summer:

```
sentinel2_rochester.npy
```

Sentinel-2 provides multispectral data across the visible, near-infrared (NIR), and shortwave-infrared (SWIR) regions. These bands come with different spatial ground sampling distances (GSD): 10 meters for B2 (Blue), B3 (Green), B4 (Red), and B8 (NIR), 20 meters for several red-edge and SWIR bands, and 60 meters for atmospheric correction bands; we will delve more into remote sensing aspect of things in the week 5. For consistency and analysis, all bands have been resampled to 30 meters. The provided dataset is surface reflectance.

Plot each of the 12 Sentinel-2 bands separately, ensuring that you identify what each band corresponds to in terms of its wavelength. Use a vibrant colormap, like `cmcocean` library. What is your approach to a proper visualization in terms of stretching (implement this in your visualization)? You can see part of the image is considered *no data*, how would you go about that? (you will use this in the following problems as well)

```
None = plot_band(args)
```

Problem 2

(a) Given the multi-spectral data for each band, define the function below that takes each band (treated as an independent random variable) and calculates the following statistics: mean, standard deviation (std), minimum, maximum, quartiles (Q1, median, Q3), as well as skewness and kurtosis which was not covered in class. Explain what each of the statistics explain. Provide the statistics for all bands and all statistics as a table.

```
stats = calculate_band_statistics(args)
```

(b) Define another function called to standardize the data for each band. This function should compute the z-scores for all pixel values in the band. Explain what standardization does to your data. Plot the histogram for each band's original data and use the standardize data to highlight the outliers. This approach helps to better understand the distribution of the data and identify any anomalies present.

```
data_standard = standardize(data)
```

Problem 3

(a) We aim to explore the relationships between variables in the multispectral data. First, use the concept of the Pearson r correlation coefficient, as discussed in class, to compute the correlation matrix for all bands. Plot this matrix as an image, where each cell represents the correlation between two bands. What type of matrix is this? Analyze the relationships between variables and describe what the correlation coefficients reveal about their linear relationships.

```
corr_matrix = correlation_matrix(args)
```

(b) For the 10-meter bands (B2, B3, B4, and B8), define a function that creates two subplots: the first displaying a pairwise scatter plot between every two vectors, and the second showing the density of scatter plot points in the scatter plot. The density plot is particularly useful for visualizing areas where data points are more concentrated (there is more than one technique to this; pick your favorite). Analyze the observed patterns and describe any significant trends or clusters in the data.

```
corr_matrix = correlation_plot(args)
```

Problem 4

Visit JPL Spectral Library. ECOSTRESS data provides high-resolution lab-based spectral reflectance across the electromagnetic spectrum, this data will be compared with the Sentinel-2 data. Download spectral data covering the **0.35–2.5 micron** wavelength range for the two materials below:

1. Oak (Quercus genus) under the vegetation category.
2. Construction Asphalt or Road under the manmade category.

Note: You can order multiple spectra at a time, but you might have to wait to receive it via email, you can also open the individual file and download the associated text file.

The spectral library data must be spectrally downsampled to the Sentinel-2 bands (Sentinel-2 bands). Use the concept of cosine similarity to identify the first 100 pixels in Sentinel-2 data that have the lowest spectral angle (the obtained angle in the cosine similarly) when compared to the ECOSTRESS samples. Plot the spectra of the 1st, 50th, and 100th closest matches alongside the original ECOSTRESS spectra. Analyze how closely the matches resemble the ECOSTRESS data and explain the similarities or differences. Pick a cut off angle and use that as a threshold to grab all the vegetation/road pixels in your sentinel-2 imagery. Represent the pixels you have identified in the image with your favorite approach.

```
angle = sam(v1, v2)
```

Notes:

- Exclude sentinel 2 bands affected by atmospheric water absorption and aerosols (443 nm, 940 nm).
- Sentinel 2 Band 10 is excluded from the data.
- Divide the ECOSTRESS reflectance data by 100 as the range is 0-100.