Metis Week 3 - Project McNulty Proposal

LandSat Imaging Data

Steven Bierer, October 16, 2018

Data Resources and Tools

This project emphasizes the use of supervised classification methods. I have chosen to analyze satellite imaging data of the earth's surface, obtained from the University of California at Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Statlog+%28Landsat+Satellite%29). The data, which was originally utilized by the Australian Centre for Remote Sensing, represent a small section of Earth's surface taken from a larger imaging database. The images date to the early 1990's and were generated by a Landsat Multi-Spectral Scanner housed on a remote sensing NASA satellite. The data are stored as numerical values in an ASCII text file.

Analysis will be performed in Python using the modules pandas, numpy, and skikit-learn. The modules matplotlib and seaborn will be used for graphical display. Other tools will be added as needed.

Description of Data Features

Each data element represents a 3x3 pixel area of a larger 82 x 100 pixel grid. Each of the 9 pixels is assigned four numbers describing the sensor magnitude in a different spectral range. Thus, there are 9 x 4 = 36 features for every row of data, and there are 6435 rows of data. In the data table, the features corresponding to a square are arranged in pixel sequence from left-to-right, top-to-bottom, with the four spectral magnitudes given consecutively for each pixel (e.g. the top-left pixel's four spectral values are in columns 1, 2, 3, and 4). Magnitudes are integers ranging from 0 = black to 255 = white.

The data were labeled manually by an observer on the ground. Each pixel (which corresponds to an 80 meter x 80 meter physical square on the ground) is labeled according to the terrain using the following code: 1 = red soil, 2 = cotton crop, 3 = grey soil, 4 = damp grey soil, 5 = soil with vegetation, 6 = mixture (all types), 7 = very damp grey soil. The “mixture” class is not labeled in this particular data set, so there will be only 6 classes.

It must be noted that the data rows do NOT come in order based on their position on the 82 x 100 grid. Moreover, some positions are missing. This was done so that the 2-d image can’t be reconstructed, such that spatial covariance information can’t be used across data elements to improve classification.

Classification Approach

I will take the following steps with the data, though depending on outcomes this workflow will need to be changed:

1) Perform unsupervised classification using the K-means algorithm (or similar) with N = 7 classes. After applying the labels, this will give me an idea of how the features cluster with respect the four spectral components. Applying the same clustering with N = 6, 5, etc, will give an indication of how “similar” the classes are to each other.
2) With the data separated into a training and test set, perform supervised classification. Logistic regression (already learned in class) would be a good start. Apply the regression with a cross-validation approach.
3) The quality of classification will be assessed by observing the different types of classification errors: Recall, Specificity, Precision, and Accuracy. If necessary, I can apply a cost function to the classification prediction, to improve performance.
4) Iterate with different classification algorithms and cost functions as needed.

Expectations

I do not anticipate major difficulties working with the imaging data for this project, as the features are all of the same type (pixel magnitudes coded as integers between 0 and 255). For some analyses, the features will have to be interpreted as a 2-D structure (i.e. not as a simple vector). Neighboring pixels and adjacent spectral components will likely be highly correlated, however such correlation can be handled by a number of proven methods, such as applying principal components to reduce the number of features or using the “Hu” image moment across pixels.