Materials Informatics – Fall 2017

Final Computer Project

Due on: Dec 15

Assignment 1: In this part you will use the SFE material data set from the previous Projects. We will use the entire data set, without splitting it into training and test data. We will no longer categorize the data into discrete classes according to SFE, but will use the continuous scale of SFE as the response. Pre-processing should be done the same way as before: discard all columns (predictors) that do not have at least 60% nonzero values; from the data that remain, remove the rows (observations) that contain any zero values. This should leave a data matrix with 211 observations and 7 predictors (C, N, Ni, Fe, Mn, Si, Cr).

1. Fit a linear regression model (with intercept) separately to each of the seven variables. List the fitted coefficients, the mean residual sum of squares, and the R^2 statistic for each model. Which one of the seven variables is the best predictor of SRE, according to R^2? Plot the SFE response against each of the seven variables, with regression lines superimposed. How do you interpret these results?

2. Perform exhaustive search (for 1 to 5 variables) and sequential forward search (for 1 to 5 variables) using the R^2 statistic as the search criterion. List the fitted coefficients, the mean residual sum of squares, the R^2 statistic, and the *adjusted* R^2 statistic for each model. Which would be the most predictive model according to adjusted R^2? How do you compare this result to the one in Project 2, using classification?

3. Now fit a linear regression model to the entire data set using ridge regression and the Lasso. In each case, use the following values for the regularization parameter $\lambda = 50, 30, 15, 7, 3, 1, 0.30, 0.10, 0.03, 0.01$. Do not apply any normalization or scaling to the data. For each case (ridge and lasso), list the regression coefficients for each value of λ, and also plot the "coefficient path."

Assignment 2:

We will use the Carnegie Mellon University Ultrahigh Carbon Steel (CMU-UHCS) dataset in

> B. DeCost, T. Francis and E. Holm, "Exploring the microstructure manifold: image texture representations applied to ultrahigh carbon steel microstructures." arXiv:1702.01117v2 (2017).

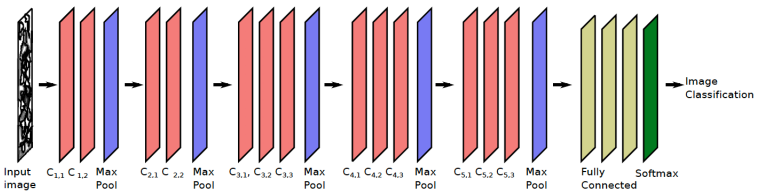
The data set is available on the TAMU Google Drive at `http://bit.ly/2jaGCkg`. There are three files: a ZIP file containing the raw images and two excel files containing the labels and sample preparation information. Please read DeCost's paper to learn more about the data set.

We will classify the micrographs according to **primary microconstituent**. There are a total of seven different labels, corresponding to different phases of steel resulting from different thermal processing (number of images in parenthesis): spheroidite (374), carbide network(212), pearlite (124), pearlite + spheroidite (107), Widmanstatten cementite (81), pearlite + Widmansttten (27), and Martensite/Bainite (36).

We will use the spherodite, network, pearlite, and Widmanstatten categories for training. The training data will be **the first 100 data points** in the spherodite, network, pearlite categories and the first 60 points in the Widmanstatten category. The remaining data points constitute the test set.

We will use a *one-vs.-one* approach to deal with the multiple labels, where each of 4 choose 2 = 6 classification problems for each pair of labels are carried out. The classification rule to be used is a Radial Basis Function (RBF) nonlinear SVM classification rule. Given a new image, each of the six classifiers is applied and then a vote is taken to achieve a consensus for the most often predicted label.

To featurize the images, we will use the pre-trained VGG16 deep convolutional neural network (CNN), which has the following architecture:



We will ignore the fully connected layers, and take the features from the intermediate layers **C1,2 C2,2 C3,3 C4,3 C5,3**, using the mean value of the filter outputs as the feature vector. In each pairwise classification experiment, we will select one of the five layer according to the best 10-fold cross-validation error estimate (with ten repetitions).

You are supposed to record the following:

(a) The convolution layer used and the cross-validated error estimate for each of the six pairwise two-label classifiers.

(b) Separate test error rates on the unused micrographs of each of the four categories, for the pairwise two-label classifiers and the multilabel one-vs-one voting classifier described previously.

(c) For the mixed pearlite + spheroidite and pearlite + Widmansttten microstructures, apply the trained one-vs-one multilabel voting classifier and find out which label it gives to each micrograph.

(d) Now apply the binary two-label classifier pearlite vs. spherodoite to the mixed pearlite + spheroidite micrographs. Repeat for the pearlite + Widmansttten microstructures. Compare to item (c).

(e) For the martensite microstructure, which was not trained for, apply the trained one-vs-one multilabel voting classifier and find out which label it gives to each micrograph.

In each case above, interpret your results. Implementation should use the Scikit-Learn and Keras python libraries.