Materials Informatics – Fall 2018

Computer Project 3

Due on: Nov 21

This project uses a data set on Fe-based nanocrystalline soft magnetic alloys, collected from different papers by Yuhao Wang from Dr. Arroyave's group (the dataset is available on e-campus). This data set records the atomic composition and processing parameters along with several different electromagnetic properties for a large number of magnetic alloys. In this project, we will use unsupervised learning to correlate the atomic composition features and the aneealing temperature processing parameter to the magnetic coercivity of the material. Larger values of coercivity mean that the magnetized material has a wider histeresis curve and can withstand larger magnetic external fields without losing its own magnetization. By constrast, small values of coercivity mean that a material can lose its magnetization quickly. Large-coercivity materials are therefore ideal to make permanent magnets, for example.

1. (Data preprocessing.) The features are the atomic composition percentages (first 25 columns of the data matrix) and the annealing temperature (column 26), while the response vatiable is the magnetic coercivity in A/m (column 34). Preprocessing consists of the following steps:

(a) read the spreadsheet into python;

(b) discard all features (columns) that do not have at least 5% nonzero values;

(c) discard all entries (rows) that do not have a recorded coercivity value (i.e., discard NaNs).

(d) add zero-mean Gaussian noise of standard deviation 2 to all feature values (but not the response variable). Do not be concerned with possible negative values. To ensure reproducible results, set numpy's random seed to zero (execute `np.random.seed(0)` prior to generating the noise).

(e) normalize all feature vectors to have zero mean and unit variance.

Note that **the order in which the operations are applied matter.** Step 4 is necessary to introduce a small degree of randomness to the data. Step 5 is necessary because the features are measured on different scales (percentage for the atomix composition and Kelvin for the annealing temperature); in addition, most of the alloys have large Fe and Si composition percentages. Without normalization, Fe, Si, and the annealing temperature would unduly dominate the analysis.
Hint: normalize the data using the `StandardScaler` function in the `sklearn.preprocessing` module.

2. Run PCA on the resulting 741×12 feature data matrix.
Hint: use the `PCA` function in the `sklearn.decomposition` module.

3. There are 12 principal components. Plot the percentage of variance explained by each PC as a function of PC number (this is called a "scree" plot. Now plot the cumulative percentage of variance explained by the PCs as a function of PC number. How many PCs are needed to explain 95% of the variance?
Hint: use the attribute `explained_variance_ratio_` and the `cusum()` method.

4. Obtain scatter plots of the first few PCs against each other (PC1 vs. PC2, PC1 vs. PC3, and PC2 vs. PC3). In order to investigate the association between the features and the coercivity, categorize the latter into three classes: "low" (coercivity $\leq$ 2 A/M), "medium" (2 A/M $<$ coercivity $<$ 8 A/M), and "high" (coercivity $\geq$ 2 A/M). Color code the previous scatter plots using red, green, and blue to identify high, middle, and low coercivity. What do you observe? If you had to project the data down into just one PC, while retaining maximum discrimination, which one would it be?

5. Print the "loading" matrix W (this is the matrix of eigenvectors, ordered by PC number from left to right). The absolute value of the coefficients indicate the relative importance of each original variable (row of W) in the correponding PC (column of W). Identify which two features contribute the most to the discriminating PC you identified in the previous item. This is an application of PCA to feature selection.

6. Run hierarchical clustering using Scipy's clustering package (`scipy.cluster`) on the two features selected in the previous step. Deterministiacally sample the data set by taking rows $0, 12, 24, \ldots, 132$. Construct dendrograms for single, average, and complete linkage using the Euclidean distance. Label the leaves of the dendrograms with "High," "Mid," and "Low" according to the coercivity. What do you conclude from this?
Hint: There should be general agreement between the different linkage methods and the presence of two main clusters and a couple of outliers.