Materials Informatics – Fall 2017

Computer Project 3

Due on: Nov 7

We will use the SFE material data set from Projects 1 and 2 (we will use the entire data set, without splitting it into training and test data). We will categorize the data into three classes, low (SFE < 35), middle ($35 \leq$ SFE ≤ 45) and high (SFE > 45) stacking fault energy. Pre-processing should be done the same way as before: discard all predictors that do not have at least 60% nonzero values; from the data that remain, remove the rows (samples) that contain any zero values.

Assignment 1: Perform dimensionality reduction by PCA.

(a) Run PCA on the 211×7 data matrix (in R, use the function `prcomp`, with the scale option set to TRUE).

(b) There are 7 principal components. Plot the percentage of variance explained by each PC as a function of PC number.

(c) Plot the first few PCs against each other, using red, green, and blue to identify high, middle, and low SF energy (you can generate a single image with all 42 pairs of plots using the R function `pairs`). What do you observe? If you had to project the data down into just one PC, which one would capture the difference in SF energy the best: the first PC, the second PC, or the third PC?

(d) Now look at the "loading" matrix W (that 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 into the correponding PC (column of W). Identify which elements contribute the most to the discriminating PC you identified in the previous item. How do you compare this to the results obtained in Projects 1 and 2?

Assignment 2: Perform hierarchical clustering.

(a) Run hierarchical clustering on the 211×7 data matrix, using single, average, and complete linkage and the Euclidean distance (you can do that using the functions `hclust` and `dist` in R. Plot the corresponding dendrograms; label the leaf nodes according to whether they correspond to low, middle, and high SF energy (you can do that in R by executing `plot` on the output of `hclust`). How do you compare the three linkage methods? What can you say about the structure of the data based on the dendrograms?

(b) Cut each dendrogram in order to obtain three clusters (you can do that in R with the funtion `cuttree`). Compare the three categories (low, middle, and high) SFE to their cluster memberships. What do you observe?