Hwk #4: Regression and Classification
Biostat 202: Opportunities and Challenges of "Big Data"

For this homework we will use the "breastcancer.csv" dataset available on the CLE. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset contains features that are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute Information: The dataset consists of an Record ID (idno) in the first column followed by image features each graded from 1 to 10: "ClumpThickness", "CellSizeUniformity", "CellShapeUniformity", "MarginalAdhesion", "SingleEpithelialCellSize", "BareNuclei", "BlanChromatin", "NormalNucleoli", "Mitoses". The final variable indicates the diagnosis "Class" as to whether the sample is "benign" or "malignant".

Clustering

1. (1 point). Read the dataset in from the course CLE. It is called "breastcancer.csv".
2. (1 point). Use the type node to set the variables appropriately. The image features should be considered as continuous variables with Role as Input. Initially, set the Class variable to have Role "None"
3. (2 points). Partition the data as Training (60%) / Test (40%)
4. (3 points). Run K-means algorithm with K = 2, 5, 8, 12.
5. (2 points). Which one of the 4 models has the lowest Silhouette score based on each of the Training and Test datasets?
6. (2 points). Why might the number of clusters in the Test set be less than the number of clusters that were fit in the Training set?
7. (4 points). Given the observed Silhouette scores, describe how might you now go about finding the single best value of K (i.e. for all possible numeric choices of K)?
8. (3 points) What is the purpose of partitioning the dataset even though there is no gold-standard outcome to test against?

Classification

9. (1 point). Edit the Type Node such that Class is set to the Target variable.
10. (2 points). Partition the data as Training (40%) / Test (30%) / Validation (30%)
11. (2 points). Fit Logistic Regression, C5.0, Neural Networks, and Bayesian Networks with the default settings and with the Target variable as "Class".
12. (2 points). Which model performs best in the Training Set and which model performs best in the Test Set in terms of overall accuracy (based on the View selector in the Model Tab)?
13. (4 points). Examine the results with an Analysis Node. What kind of accuracy do we see in the Bayesian Network model for, a) the Training set, b) the Testing set, and c) the Validation set? (Note that I am using IBM Modeler specification of training/test/validation.) Do the same accuracy comparison for the Neural Network. Briefly comment on the differences in results.