CMSC 435 Assignment 3

Fall 2017
(individual work; 10 pts total)

The assignment asks you to compute, evaluate and compare models for the prediction of protein crystallization using a provided dataset.

Dataset
The dataset ("dataset_assignment3.csv" file) is provided in the text-based, comma-separated format where each protein is represented by 84 numeric features and 1 symbolic outcome:

- Composition of amino acids: AAcomp_{AA} (20 features)
- Physiochemical properties of proteins: AAindex_{name}_avg (64 features)
- Outcome (1 features), which is encoded as: F (failed to crystallize) and C (crystallizable)

The dataset includes 1204 crystallizable and 2383 proteins that failed to crystallize

Development of predictive models
You are required to compute models with version 7.6 of the RapidMiner Studio using five different algorithms. Three of these algorithms must be the Decision Tree, Logistic Regression and Naïve Bayes. You can choose any of the other classification algorithms for the remaining two. You should parametrize each of these algorithms, i.e., select the best possible combination of values of their (key) parameters, to the best of your ability, to maximize predictive performance. Note that you will need to read, make an educated guess, and/or use trial-and-error approach to figure out which parameters make a difference and how to use them. Do not attempt to reduce the size of the dataset, i.e., do not perform feature or sample selection.

Evaluation and comparison of predictive models
You must evaluate the predictive performance using accuracy ("% of correctly classified instances"). For each method you must perform three types of tests:

- on the entire dataset ("use training dataset")
- on 50% of the dataset; you will use the other 50% to compute the model ("percentage split")
- using the 10 fold cross-validation

The 10 fold cross-validation divides the dataset at random into 10 equal-size subsets, where one subset is used to test the model and the remaining nine to compute the prediction model. This is repeated 10 times, each time using a different subset as the test set. Consequently, this test results in predicting every protein in the dataset. This test type is implemented in the RapidMiner Studio with the "Cross Validation" operator where the number of folds is set to 10.

Deliverables

1. List and briefly describe the methods that you used and list their key parameters.
2. Using a table (see below), report the accuracies for the five algorithms and the three test types. You must include the accuracies of the models that use default parameters and the best selected parameters. In total, you have 5*3*2 = 30 results to report. List the best selected values of parameters for each model and each test type.

3. Discuss which of the three types of the tests would be appropriate to give the most reliable estimate of predictive performance, i.e., the performance that a user of your model should expect to observe on new proteins that were not included in the provided dataset.
4. Discuss whether trying multiple algorithms and adjusting their parameters helped you in developing a more accurate predictive model and if yes then whether the corresponding amount of the improvement justifies the amount of effort.
5. Discuss whether the accuracy of your most accurate model is sufficient for practical purposes. Justify your answer.
6. Provide the "confusion matrix" that corresponds to your best result (out of the 30) and use it to explain whether this model would be suitable to identify proteins that crystallize, proteins that do not crystallize, or both types of proteins.

NOTES

– The table mentioned under the second deliverable must be in the following format; for your convenience this table is provided in the word doc format on the Blackboard.

Reported information	Test type	Decision Tree	Logistic Regression	Naïve Bayes		
Accuracy with default parameters	Entire dataset					
	50%					
	Cross-validation					
Accuracy with best parameters	Entire dataset					
	50%					
	Cross-validation					
List names of parameters						
List selected best values of parameters (in the same order as in the list of names)	Entire dataset					
	50%					
	Cross-validation					

– Use a separate, **clearly marked paragraph** for each of the six deliverables.

Due Date

Your assignment must be received by 13:45pm Eastern Time, October 10 (Tuesday), 2017. It should be typed single-spaced, using 12 point font size and with standard margins. Only hardcopies will be accepted by the end of the class.