CMSC 435 Assignment 3

Fall 2019
(individual work; 10 pts total)

This assignment asks you to develop, evaluate and compare models for the prediction of proteins that interact with nucleic acids using a provided dataset.

Dataset

The dataset (*dataset_a3.csv* file) is provided in the text-based, comma-separated format where each protein is represented by 8 numeric features and 1 symbolic outcome. The outcome feature (called "Class") annotates each proteins as *Yes* (interacting with nucleic acids) vs. *No* (non-interacting). The dataset includes 8795 proteins, with 936 labeled *Yes* and 7859 labeled *No*.

Development of predictive models

You are required to compute models with version 9.3 (or higher) of the RapidMiner Studio using four different algorithms. Three of these four algorithms must be the Decision Tree, SVM and Naïve Bayes. You can choose any of the other predictive algorithms for the fourth methods. You should parametrize each of these algorithms (select the best possible combination of values of their parameters), to the best of your ability, in order 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 use the "advanced parameters"**. Do not attempt to sample the dataset, i.e., do not perform feature or sample/object selection.

Evaluation and comparison of predictive models

You must evaluate the predictive performance using accuracy ("% of correctly classified instances"). For each algorithm 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 5 fold cross-validation

The 5 fold cross-validation divides the dataset at random into 5 equal-size subsets, where one subset is used to test the model and the remaining nine to compute the prediction model. This is repeated 5 times, each time using a different subset as the test set. Consequently, this 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 5.

Deliverables

1. **List and briefly describe** the methods that you used (one sentence per method). Provide a **list key parameters** for each method, i.e., parameters that allowed you to improve results when compared with the default parameter values. The key parameters could/should be a subset of all available parameters.
2. Using the table shown below, **report the accuracies** for the four algorithms and the three test types. The accuracy values must be reported with two digits after the decimal point, e.g., 91.05. You must include the accuracies of the models that use the default parameters and the

best selected parameters. In total, you have 4*3*2 = 24 results to report. **List the best selected values of parameters** for each model and each test type.

3. **Briefly explain** 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. If yes then **comment** on whether the corresponding amount of the improvement justifies the amount of effort. Make sure that you rely the **most appropriate test results** (see question 3) when answering this question.
5. **Discuss** whether the accuracy of your most accurate model is sufficient for practical purposes. **Justify** your answer.
6. **Give** "confusion matrix" for the most accurate result computed based on the cross validation experiments (selected among the 8 corresponding experiments). Use this matrix to **explain** whether this predictor would be suitable to identify proteins that interact with nucleic acids (Class = *Yes*), proteins that do not interact with nucleic acids (Class = *No*), or both types of proteins.

NOTES

– Use a separate, **clearly marked paragraph** for each of the six deliverables.
– The table from the second deliverable must be in the following format; for your convenience this table is provided in the word docx format on the Blackboard. Example values are in green.

Reported information	Test type	Decision Tree	SVM	Naïve Bayes	
Accuracy with default parameters	Entire dataset	12.34			
	50%	23.45			
	Cross-validation	34.56			
Accuracy with best parameters	Entire dataset	45.67			
	50%	56.78			
	Cross-validation	67.89			
List names of parameters		maximal depth apply pruning … criterion			
List selected best values of parameters (in the same order as in the list of names)	Entire dataset	10 True … gain_ratio			
	50%	13 True … gain_ratio			
	Cross-validation	-1 False … informaton_gain			

Due Date

Your assignment must be received by 12:45pm Eastern Time, October 3 (Thursday), 2019. It should be typed single-spaced, using 12 point font size and with standard margins. Only hardcopies will be accepted at the beginning of the class.