COSC74/174: Machine Learning and Statistical Data Analysis

Homework 3 (due: 10AM on Tuesday, 12 February 2019)

Instructor: Prof. V.S. Subrahmanian (vs@dartmouth.edu)

You are given a new training dataset with 5,600 rows:

- Columns 1 through 6 of the given CSV file represent independent variables (IVs)
- The last column ("Label") represents the dependent variable (0 or 1)

You are required to learn a Support Vector Machine (SVM) for this project using this training data. Your classifier will be evaluated on a test set with 2,400 rows for which the last column is blank (i.e. you do not know the true class to which rows in the test data belong).

TASK

Use an SVM classifier, with hyperparameter optimization and 10-fold cross validation, to predict the class label in the test set.

Please include the following in your Canvas submission as *two separate files*:

1. A new CSV file of the test set with an added column, "Label", showing the dependent variable (0 or 1) that you predicted
2. Your python code (which can be either a Jupyter Notebook or python script)

GRADING CRITERIA

- *Accuracy on the test set (15 points)*. Your score will be your test-set accuracy divided by the benchmark accuracy (i.e. the maximum obtained by all students who submitted the assignment plus the TAs').
- *SVM hyperparameter optimization and 10-fold cross validation (3 points)*. You need to show evidence in your code that you optimized SVM hyperparameters, which should include penalty parameter C, kernel coefficient λ, and different kernels. Optimizing additional hyperparameters, if applicable, is highly encouraged, as doing so may help improve your classifier's performance. Please refer to the *scikit-learn* documentation (https://scikit-learn.org/stable/modules/svm.html) to learn about the appropriate ranges or options of these hyperparameters.
- *Python code and output (2 points)*. Your python code can be run error-free, implemented SVM, and is well-organized. Your output CSV file is properly formatted as described above.
- *Late-submission policy*. All projects will be due by the deadline on Canvas. If your submission is up to 1 day late, you will only get 80% of your original score (e.g. if your scored 18/20, you will get 0.8*18=14.4). If your submission is up to 2 days late, you will get 60% of the points you scored on that part. If your submission is 2 or more days late, your homework will not be accepted.

ADDITIONAL NOTES

- You should abstract the training and prediction phases as distinct functions:
    - ***train(data)***, where input data is the CSV filename of the training dataset, and output is a classifier *C*.
    - ***predict(C, row)***, where input *C* is a classifier object on *train(...)* was called, input is an array corresponding to IVs in the CSV test set. The output should be 0 or 1.
- You may choose to add, drop, or preprocess/transform features as in Homework 1.
- The project *must be* implemented in python. You are responsible for making sure that your project is properly submitted and your code can be properly run.
- You may use functions in the scikit-learn library (http://scikit-learn.org/stable/install.html). If you used or referenced code from the Internet that is not part of a standard python library, you must provide appropriate citation(s).
- Please be sure to submit all parts of your homework in the required format. Points will be deducted if your code is poorly organized or cannot be run error-free, or the output is improperly formatted.
- All work must be your own. Academic Honor Principle applies to all parts of the project. Please refer to http://student-affairs.dartmouth.edu/policy/academic-honor-principle for more detail and ask your instructor/TAs for clarifications.