MATH 2960 Intro to Data Mathematics Spring 2016
Classification Mini-Project

The chemical company, Chems-R-Us, has hired your company to consult on data analytics. Chems-R-Us' goal is to create a model to predict ready biodegradation of chemicals by using molecular descriptors. A data set of with molecular descriptors and biodegradation experimental values of 1055 chemicals were collected from the webpage of the National Institute of Technology and Evaluation of Japan (NITE). The company wants your consulting to develop a computational model to predict the ready biodegradability of molecules. To accomplish this they have provided the data file DatasetA.csv. Each line of the dataset consists of a molecule id, 41 attributes describing the molecule, and whether the molecule is biodegradable or not in experiments. The first column is an ID number. The 2nd through 42 columns are the attributes, and the last column is 1 if the compound is readily biodegradable and -1 if it is not. The names of each column are in the file DatasetANames.csv In addition, they have provided DatasetV which contains molecules described by the 41 attributes but for which the biodegradability has not been provided. DatasetV.csv the same format except the class column is missing.

On March 30, each company should turn-in a paper copy of project report, upload Matlab scripts/files and DatasetV prediction file to the LMS Classification Project Assignment, and give 8 minute project presentation. Details are as follows.

1. Form your consulting company. Each consulting company must consist of 2 to 3 students. You should submit one hard copy project report per company. Provide the name of your company and the names of the students involved in your project report.

2. Produce a clearly-written grammatically correct report which includes the following items

   (a) An introduction giving an overview of the report.

   (b) A basic description of the data. Describe the size of the data (number of attributes, number of points in each class). Provide the mean and covariances for each class in DatasetA. Provide the mean and covariance of all of datasetV. (Hint: the Matlab imagesc command can be a good way to display covariances as images). Describe any observations that you have.

   (c) Show how Fisher compares with the Mean Method. The company is considering saving money by just assigning points to their closest mean of each class. Show them that your company can do better. Do a study comparing Fisher LDA versus the Mean method where 90% of DatasetA is used for training and 10% for testing. Include in your report the normal and threshold of the separating planes from each method, analysis of the training errors, analysis of the testing errors, and a discussion of which method is preferable.

   (d) Describe the predictive model you suggest for predicting ready biodegradability of new compounds. This could be one of the models that you have made or a new one that you think would be better. Describe the process you use to make the model. Specify the model in full detail. For Fisher or other linear models this is done by specifying the hyperplane normal and threshold. If you do some other type of model, please provide an appropriate equation or Matlab code for the final model.

   (e) Report how well your model does on DatasetA in terms of class 1 error, class -1 error, and total error.

   (f) Report how well you estimate your model will do on future data. Describe your procedure for estimating this. Give your estimates of class 1 error, class -1 error, and total error.

   (g) Report how many points of DatasetV are estimated by your model to be Class 1 and how many are estimated to be Class -1.

   (h) Provide a conclusion which summarizes your results briefly and adds any observations/suggestions that you have for Chems-R-Us about the data, model, or future work.

(i) Optional extra credit, provide any additional analysis or visualizations that may be insightful to Chems-R-Us or any extra steps you came up for improving your final predictive model (use your imagination, extra credit for creativity here).

3. Provide a csv file with your predictions of the biodegradibility of DatasetV. Make the file name contain the name of your team. Chems-R-Us will use this to independently verify the quality of your results. These predictions should be given as a csv files with the id number and prediction (1 or -1) for each points in DatasetV. Upload your final prediction to the Classification Project Assignment in the LMS account of one team. Indicate in your report which team member this is. HINT you can use the csvwrite command to write a matrix to a csv file. The syntax to write matrix M to a csv file named foo.csv is `csvwrite('foo.csv', M)`.

4. Provide one or more published Matlabs scripts that execute all the results that you gave in your report. If you write additional matlab or other code, please submit that code as well. Upload these scripts to the LMS Classification Project Assignment for the designated team member.

5. Each team should give a eight minute oral presentation summarizing your results. Be prepared to give it in class on the due date.

The company will evaluate your teams work using the following criteria

1. (30 pts) Was item 2.c, the study of the Fisher LDA and the Mean Method with the 90% train and 10% test set successfully done?

2. (10 pts) Was the procedure for constructing the final predictive model well thought-out, described, and executed?

3. (10 pts) Was the procedure for evaluating the generalization of the final model well thought-out, described, and executed?

4. (20 pts) Were all of the remaining items in 2.a-h. above included?

5. (10 pts) What is the grammatical quality and clarity of the written report? Did you communicate your results effectively in written form.

6. (10 pts) How accurate is the prediction of Dataset V as compared with the ground truth? You'll get full credit if you are close to the best results in the class.

7. (10 pts) The effectiveness of your six to eight minute pitch of your methods and results at conveying your message. Things to think about: ability to convey message to audience, quality of visual aids, visual aids prepared in advance, clear delivery, and well rehearsed.

8. (Extra credit up to 5 pts) Extra Credit will be given for creativity. So feel free to experiment with how you produce the final model and to include additional analysis that may be helpful to Chems-R-Us or that supports your model.