CMSC 435 Project

Fall 2017
(Group work; 15 pts for report + 10 pts for presentation = 25 pts total)

The project asks you to develop, evaluate and compare models for the prediction of protein crystallization, production, and purification using a provided dataset. Although each group will solve the same prediction task the corresponding designs should be unique, i.e. collaboration between groups is not allowed. Your model must classify a protein, which is represented by a set of 940 features, into one of the four outcomes, i.e., material production failed, purification failed, crystallization failed, and crystallizable.

Datasets

Two datasets are/will be provided:

- *training.csv* (*training dataset*) that includes 1242 material production failed, 716 purification failed, 425 crystallization failed, and 1204 crystallizable proteins (total of 3587 proteins).
- *test.csv* (*blind independent test dataset*) that includes 3584 proteins. This is an independent test set, which means that entire design procedure (including feature selection, parameterization of your classifier, etc.) should be completed using only the training dataset. The test dataset should be used to evaluate your system only once. This dataset will be posted on the class web site 4 days before the project submission deadline and it will **not** include the annotation of the outcomes. You will have to predict the outcomes and the instructor will process and assess these predictions.

The training dataset is provided in the text-based, comma-separated format where each protein is represented by 941 numeric features including:

- Composition of amino acids: AAcomp_{AA} and AAcomp_{AA}_{AA} (420 features)
- Physiochemical properties of proteins: AAindex_{name} (448 features)
- Other properties of proteins (4 features)
- Intrinsic disorder and sequence complexity (68 features)
- Outcome (1 features), which is encoded as: *zero* (material production failed), *one* (purification failed), *two* (crystallization failed), and *three* (crystallizable)

The test dataset is in the same format as the training dataset, except that the outcome is not provided. The features are described in greater detail in the DescriptionOfFeatures.pdf file.

Evaluation of Predictions

You are required to perform the 5-fold cross validation tests when using the *training dataset*. This cross validation divides the training dataset into 5 random, equal-size subsets, where one subset is used to test the prediction model and the remaining four to train/develop the prediction model; this is repeated 5 times, each time using a different subset as the test set. Consequently, this test results in predicting every sequence in the training dataset. This test procedure is supported by RapidMiner.

For each of the four outcomes you will convert the dataset into a binary problem, i.e., a given outcome (positive outcome) vs. all other outcomes (negative outcomes). For example, all proteins that are labeled as “material production failed” will be considered as positive, and the

remaining proteins (purification failed, crystallization failed, and crystallizable) as negative. Next, for each of the four outcomes you will compute the following measures

Sensitivity = SENS = 100*TP / (TP + FN)
Specificity = SPEC = 100*TN / (TN + FP)
PredictiveACC = 100* (TP+TN) / (TP+FP+TN+FN)
MCC = (TP*TN – FP*FN) / sqrt[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]

where TP is the number of true positives (correctly predicted positive outcomes), FP denotes false positives (negative outcomes that were predicted as positives), TN denotes true negatives (correctly predicted negative outcomes), FN stands for false negatives (positive outcomes that were predicted as negatives). You will also compute:

averageMCC = ($MCC_{\text{material production failed}}$ + $MCC_{\text{purification failed}}$ + $MCC_{\text{crystallization failed}}$ + $MCC_{\text{crystallizable}}$)/4
accuracy = 100*TP_{all} / (number of all protein in the dataset)

where $MCC_{\text{material production failed}}$, $MCC_{\text{purification failed}}$, $MCC_{\text{crystallization failed}}$, and $MCC_{\text{crystallizable}}$ denote the MCC values when using the material production failed, purification failed, crystallization failed, and crystallizable outcomes as the positives, TP_{all} is the number of correctly predicted outcomes (actual material production failed proteins predicted as material production failed proteins and actual purification failed proteins predicted as purification failed proteins, etc.).

The above measures can be computed based on the confusion matrix which is provided by RapidMiner. You should **round the values** to one digit after the decimal point when reporting the accuracy, sensitivities, and specificity and to three digits after the decimal point when reporting MCC. You report should also include the confusion matrix for your final solution.

You also must provide and summarize predictions on the *blind test dataset*. To do that you will compute your model using the entire training dataset (using the same design, i.e., futures, values of parameters, etc., as in your best 5 fold cross validation result) and you will use this model to predict sequences from the blind test dataset. In your report, you must discuss the corresponding results on both the training and blind test dataset; on the blind test dataset you can summarize your results by explaining and comparing how many proteins were predicted with a given outcome.

Design

You need to **design** your predictive model to maximize its predictive performance **evaluated based on averageMCC using the 5 fold cross validation on the training dataset**. The design may consider

- Selection of a subset of the input features. This could potentially speed up computation of the model, remove weak/noisy features, and reduce overfitting.
  IMPORTANT: ensure that you perform the feature selection (and all other design activities) using the 5-fold cross validation on the training dataset. Otherwise you could overfit this dataset and your results on the test dataset could suffer. Feel free to combine results of multiple feature selection methods.
- Selection of the classification algorithm that you will use to compute your model from among many algorithms that are available in RapidMiner.
- Parametrization of the selected classification algorithm(s). This involves setting values of their key parameters.
- Consideration of how to perform the prediction. There are at least two alternatives: use one model to predict all 4 classes vs. use 4 models to predict each of the four classes. In

the latter case, you will have to combine the four results to select one "best" result for each protein. The advantage of the second approach is that you can choose different subsets of features and different classification algorithms and their parameters for each class. Consider also the fact that these four outcomes are sequential in nature, i.e., material production happens before purification, which in turn is done before crystallization.

- Build a system that consists of multiple models that are used together. For instance, you could use multiple models that predict all 4 classes and combine their results together to generate one prediction. Check the methods in RapidMiner which you find at Operators → Modeling → Predictive → Ensembles.

IMPORTANT: In your report, you should clearly indicate **one** best set of results, which must be selected based on the cross validation on the training datasets. Moreover, these results should be compared with your intermediate results (earlier designs, other alternatives, etc.) and with results of the existing method, see Table 1, to justify your design choices. **In your write up, report your results by adding them into Table 1 to make it easy to compare the various results; indicate which result is the best/final**. You are not expected to outperform the results from Table 1 but you should provide a convincing argument why and how your method is good/competitive.

Table 1. Results of the existing PPCpred method (this table is provided in the Blackboard).

Outcome	Quality measure	PPCpred based on 5-fold cross validation on training dataset	PPCpred on the blind test dataset
material production failed	*Sensitivity*	78.2	78.0
	Specificity	67.6	69.2
	PredictiveACC	74.5	75.0
	MCC	**0.449**	**0.462**
purification failed	*Sensitivity*	90.1	89.0
	Specificity	26.8	30.8
	PredictiveACC	77.5	77.4
	MCC	**0.199**	**0.222**
crystallization failed	*Sensitivity*	86.2	87.4
	Specificity	46.6	41.7
	PredictiveACC	81.5	82.0
	MCC	**0.277**	**0.257**
crystallizable	*Sensitivity*	61.0	61.2
	Specificity	83.6	84.8
	PredictiveACC	76.1	76.8
	MCC	**0.455**	**0.471**
averageMCC		**0.345**	**0.353**
accuracy		54.8	55.6

Deliverables

Each group shall provide the following three deliverables:

1. **Report** that consists of:
   - **Cover page** that gives the class number and title, date of your submission, name of your group and names of all teammates.
   - **Description of the design of the prediction system**. You should explain which algorithms and their parameters you tried and why and which you have chosen; how and whether you performed feature selection, and if yes then how many and which features were selected; and which other design options you considered and applied.

- **Results** (see *Evaluation of Predictions* section). You must organize the results in a table using the format of Table 1. Using this format, compare your best results with the results from earlier/alternative designs and with the results shown in Table 1.
- **Conclusions**. This is a **very important part** of your report. You should comment on the quality of your results and put them into perspective against the results from Table 1. Also, describe your experience in this project and explain advantages and disadvantages of your method and why you think your results are good or bad.

2. **Predictions on the *blind test* dataset**. These predictions should be submitted via email to the TA at wangc27@vcu.edu as a text file named with the mane of your group, where each row provides prediction for a given "blind" protein. The format should be as follows:

   zero
   two
   three
   one
   …

   where zero, one, two and three are the predicted outcomes for the protein from the same row in the *test.csv* file. The TA will use these results to evaluate your method on the blind test dataset and these results will be forwarded to you as part of the evaluation of your project.

3. **Presentation** that
   - is 8 minutes long plus 2 minutes for questions
   - shall describe the design, results and conclusions
   - shall include the following parts:
     - Motivation for your design. You will need to explain how you arrived at your final design.
     - Discussion and comparison of the quality of the achieved best results using the results on the training dataset and Table 1.
     - Conclusions. This part is essential; see the conclusions part of your report.

Marking

The evaluation of the project report and predictions constitutes 15% of the final mark from the course and it will consist of the following three parts:

1. 30% for the quality of the report
2. 20% for the quality of the design of the prediction method
3. 50% for the quality of the predictions measured using the 5 fold cross validation on the training dataset and on the blind test set.

NOTES: For item 3, the *averageMCC* is of the primary importance. This is the quality measure that will be use to rank and evaluate all submitted solutions, but the conclusions **must discuss** the other quality indices as well. MCC that is high(er) relative to other submissions or results in Table 1 is not necessary to receive a full mark, i.e., it is important to discuss how is the MCC value of your best design high relative to your own alternative solutions and what advantages are provided by your method, say compared to the results in Table 1. **Bonuses** of 15%, 10%, and 5% will be given to the project submission that obtains the highest, the second highest and the third highest average value of *averageMCC* on the training and blind test datasets. In case of a tie the winner will be decided based on the higher value of the average of *accuracy*.

The presentation constitutes 10% of the final mark from the course and will be evaluated by the instructor, TA and your peers. The grade will consist of three parts:

1. Grade assigned by the fellow students (**30%**). Each project group will complete a short evaluation form, see appendix A, to assess presentations of other groups. Instructor will gather and process these grades; they will be kept confidential. We strongly advise you to reassess and potentially revise your scores after all presentations on a given date are *completed* to assure **consistency**.
2. TA's grade (**30%**). TA will grade the quality of presentations using form in Appendix B.
3. Instructor's grade (**40%**). Instructor will grade the quality of presentations using form in Appendix C.

Your overall mark, broken into the mark from peers, TA and instructor and including comments will be released to you at the final exam.

Deadlines and Delivery

Submission of the reports and the predictions is due on November 30 (Thursday), 2017, before 13:45pm. The report should be delivered as a hard copy in the classroom and predictions should be send by email to wangc27@vcu.edu.

The presentations will be delivered on December 5 (Tuesday) and 7 (Thursday), 2017, at 12:30pm (during the last two lectures); each date will include five to six groups. The schedule will be posted on the blackboard at least two weeks in advance of the presentations. Each presentation must be submitted electronically via email entitled "CMSC 435 presentation" in the PPT or PDF format to the instructor at lkurgan@vcu.edu **at least 24 hours in advance** of the corresponding presentation time. The instructor will acknowledge receiving the presentations via reply email and bring them on his laptop to the presentation session.

Final Notes

- Failing to deliver the presentation will result in 0 marks for the presentation. The will be no make-up presentations.
- Always copy the email communications to yourself so you can prove that it was sent.
- Do not cheat (e.g., do not inflate or "tweak" the results). It is better to report honest results than to get caught cheating. In the latter case you are risking receiving 0 marks for the project.
- Please contact the instructor immediately if any problems occur.

Appendix A

CMSC 435 Intro to Data Science
Fall 2017
Peer Evaluation Form for Project Presentations

Date (circle the correct date) December 5, 2017 or December 7, 2017

Name of the presenting group ……..……

Remarks:

- For each question enter grade between 0 and **10** (0 being the worst, 10 being the best)
- Optionally please add comments (both positive and negative)
- Average of these grades across all groups will be used to come up with the 30% of the peer-evaluation component.

remarks	***grade***
Quality of Presentation (Did you find the presentation interesting? Were the presenters prepared? Did you understand the topics covered in the presentation? How much did you learn? Was there anything significant missing? Were the conclusions and discussion of results covered sufficiently? How would you rate handling the discussion/questions?)	min 0, max **10** ……
Presentation Style (Quality of presentation style – Was it finished on time? Too fast/slow? Well presented? Was the presenter just reading the slides or was (s)he presenting the material beyond the content of the slides? Was there an eye contact?)	min 0, max **10** ……
Quality of Slides (Quality of slides – Did you find the slides too crowded? Too brief? Too many? Easy to read? Was the layout of individual slides appropriate and consistent? How was the overall quality of the organization, in terms of the order and flow of the slides?)	min 0, max **10** ……
Additional Comments	

Appendix B

CMSC 435 Intro to Data Science
Fall 2017
TA's Evaluation Form for Project Presentations

Date (circle the correct date) December 5, 2017 or December 7, 2017

Name of the presenting group ……..……

TASK	***grade***	*max grade*
Quality of *Motivation for the proposed design*		5
Quality of the *Discussion and comparison of the quality*		5
Quality of *Conclusions*		10
Quality of the Presentation and Presentation Style		10
TA's total mark		**30**

Appendix C

CMSC 435 Intro to Data Science
Fall 2017
Instructor's Evaluation Form for Project Presentations

Date (circle the correct date) December 5, 2017 or December 7, 2017

Name of the presenting group ……..……

TASK	***comments***	***grade***	*max grade*
Submission of presentation on time	(up to -2 points penalty)		Y / 0
Presentation finished on time	(up to -2 points penalty)		Y / 0
Quality of *Motivation for the proposed design*			5
Quality of the *Discussion and comparison of the quality*			10
Quality of *Conclusions*			10
Quality of the Presentation and Presentation Style			15
Instructor's total mark			**40**