CMSC 435 Project

Fall 2020
(Group work; 35 pts total)

The project asks your project group to develop, evaluate and compare models for the prediction of proteins that interact with DNA and RNA using a provided dataset. Your model must classify a given protein sequence into one of four outcomes, i.e., interacts with DNA (DNA), interacts with RNA (RNA), interacts with both DNA and RNA (DRNA), and does not interact with DNA or RNA (nonDRNA). Although each group will solve the same task, the corresponding designs must be unique, i.e., collaboration between groups is not allowed.

Datasets

Two datasets are/will be provided:

- *sequences_training.txt* (*training dataset*) that includes 391 DNA proteins, 523 RNA proteins, 22 DRNA proteins, and 7859 nonDRNA proteins, for the total of 8795 proteins.
- *sequences_test.txt* (*blind test dataset*) that includes 8794 proteins, with similar proportions between the four classes of proteins. This is an independent test set, which means that entire design procedure (including feature generation, feature selection, parameterization and selection of classifiers, 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 2 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 comma-separated format where each protein is represented by:

- the amino acid sequence
- the class encoded as DNA, RNA, DRNA, and nonDRNA

Test dataset will be the same format as the training dataset, except that the outcomes will not be provided.

Evaluation of Predictions

Your group is required to perform the 5-fold cross validation 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 your group 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 DNA will be considered as positive, and the remaining proteins (RNA, DRNA and nonDRNA) 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 = $(\mathrm{MCC}_{\mathrm{DNA}} + \mathrm{MCC}_{\mathrm{RNA}} + \mathrm{MCC}_{\mathrm{DRNA}} + \mathrm{MCC}_{\mathrm{nonDRNA}})/4$
accuracy = $100*\mathrm{TP}_{\mathrm{all}}$ / (number of all protein in the dataset)

where MCC_{DNA}, MCC_{RNA}, MCC_{DRNA}, and $MCC_{nonDRNA}$ denote the MCC values when using the DNA, RNA, DRNA, and nonDRNA outcomes as the positives, and TP_{all} is the number of correctly predicted outcomes (DNA proteins predicted as DNA proteins, RNA proteins predicted as RNA proteins, etc.). These measures can be computed based on the confusion matrix. 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 must include the confusion matrix for your final/best solution**.

Your group must also 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., features, 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

Your group should **design** the model to maximize its predictive performance **evaluated based on averageMCC using the 5-fold cross validation on the training dataset**. The design may consider:

- Use of different features to encode the input protein sequence. The data mining algorithms require a rectangular dataset with a fixed size and structure of the feature vector for each object (protein). Thus, you will need to convert the input protein sequences (that have variable length) into a fixed set of (numerical) features. Lecture set 7 includes a few suggestions.
- Selection of a subset of the input features. This could potentially speed up computation of the model, remove weak/noisy features, and reduce overfitting. 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.
- Building a system with 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 at Operators → Modeling → Predictive → Ensembles.
- Different ways 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 outcome/class.

NOTE 1: Ensure that your team perform all design activities (e.g., feature selection, selection and parametrization of the classification algorithms, etc.) using the 5-fold cross validation on the training dataset. Otherwise you could overfit this dataset and your results on the blind test dataset could suffer.

NOTE 2: Your team’s design should be done incrementally. Start with a simple initial solution (complete the entire design, prediction, and prediction assessment process) and gradually make your design more sophisticated with the goal to improve the predictive performance. The progress report checkpoint should be based on a simple initial solution. In your final report, you should clearly indicate **one** best set of results, which must be selected based on the cross-validation results on the training dataset. Moreover, these results should be compared with your intermediate results (earlier/simpler designs, other alternatives, etc.) and with baseline results shown in Table 1, in order to justify your design choices. **In your final report, provide your results by adding them into Table 1.** This will make it easy to compare the different alternatives. **Clearly indicate which result is the best/final**. You should explain how you made decisions that led you a certain direction of redesigning your model. You also should provide a convincing argument why and how your method is good/competitive in comparison to the baseline result in Table 1.

Table 1. Predictive results based on the 5-fold cross validation on the training dataset (this table is available in the Blackboard).

Outcome	Quality measure	Baseline result	Design 1	Design 2	Design 3	Best Design
DNA	*Sensitivity*	6.9				
	Specificity	99.3				
	PredictiveACC	95.2				
	MCC	**0.132**				
RNA	*Sensitivity*	39.6				
	Specificity	98.9				
	PredictiveACC	95.3				
	MCC	**0.501**				
DRNA	*Sensitivity*	4.5				
	Specificity	100.0				
	PredictiveACC	99.7				
	MCC	**0.122**				
nonDRNA	*Sensitivity*	98.6				
	Specificity	29.8				
	PredictiveACC	91.3				
	MCC	**0.428**				
averageMCC		**0.296**				
accuracy		90.8				

Deliverables

Each group shall provide the following five deliverables:

1. **Progress 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 team members.
   - **Description of the first attempt to make predictions**. You should use short bullet points to list the features that you generated from the input sequences; list major data processing steps that you used to prepare the data; and name the classification algorithm that you used. This is supposed to be an initial attempt so we expect to see simple models and just a few bullet points.
   - **4x4 confusion matrix and the four MCC values** (MCC_{DNA} + MCC_{RNA} + MCC_{DRNA} + $\text{MCC}_{\text{nonDRNA}}$) that the above model has produced.
   - **Training dataset file** in txt/csv format. This file is the rectangular dataset with a fixed set of features for each object (protein) where the last (right-most) feature is the class. This is supposed to be an initial version of the dataset and so the feature set should be small and simple.
2. **Final 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 team members.
   - **Description of the design of the prediction system**. You should briefly explain the features that you generated from the input sequences; how and which features were selected; which classification algorithms and their parameters you tried and why and which you have chosen; 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 cross validation results with the results from earlier/alternative designs and with the results shown in Table 1. Include confusion matrix for your best solution. Summarize predictions for the *blind test dataset*.

- **Conclusions**. This is a **very important part** of your report. You should comment on the quality of your results and compare them against the baseline 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, in comparison with the other results from Table 1.

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

   DNA
   DNA
   RNA
   nonDRNA
   …

   where DNA, RNA, DRNA and nonDRNA are the predicted outcomes for the protein from the same row in the *sequences_test.txt* file. The instructor will use these results to evaluate your method on the blind test dataset against the true classes, and these results will be forwarded to you as part of the evaluation of your project.
4. **Video presentation**
   - Up to **6 minutes long** pre-recorded video presentation. The presentation can be done using Zoom recording to a local computer. This should include screen share of the presentation with the video of the presenter(s).
   - shall describe the design, results and conclusions
   - shall include the following parts:
     - Motivation for your design. Briefly explain how you arrived at your final design.
     - Description of your design. Explain (preferably with a diagram) how your method makes the predictions.
     - 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.
5. **Statement of contributions**
   - A short document with bullet-point style list of detailed contributions to the project for each team member. The contributions cover all aspects of the project including conceptualization and design of the methodology, implementation, testing, writing the report, preparing the presentation, making the presentation, coordination of the work, notes taking, etc.
   - The contribution list for each team member should be accompanied with an estimated fraction of the total project effort, quantified in %. The effort estimates across the team members must sum up to 100%. Each team should strive to balance the effort to be equal across team member.
   - This statement will be used to distribute the project grade among the team members.

Marking

The evaluation of the project report and predictions constitutes **20% of the final mark from the course** and it will consist of the following three parts:

1. 5% for the quality of the progress report
2. 5% for the quality of the final report
3. 4% for the quality of the design of the prediction method from the final report
4. 6% for the quality of the predictions measured using the 5-fold cross validation on the training dataset from the final report and on the blind test dataset

NOTE 3: For item 4, the *averageMCC* is the main predictive quality measure that will be used to evaluate submitted solutions but the conclusions **must discuss** the other quality indices as well. **Bonuses** of 3%, 2%, and 1% will be given to the project submissions that secure the highest, the second highest and the

third highest value of *averageMCC* on the blind test dataset. In case of a tie the winner will be decided based on the higher value of the *accuracy* on the blind test dataset.
NOTE 4: For item 4, MCCs that are high(er) relative to other submissions or to the baseline in Table 1 are not necessary to receive a full mark. The key aspect is to show substantial progress from the initial solution – you should show and discuss how your best design is better when compared to your own alternative solutions and explain advantages compared to the baseline results in Table 1.

The video presentation constitutes **15% 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 (peer evaluation) (**5%**). Each project group will complete a short evaluation form online, see appendix A, to assess presentations of other groups. Instructor will gather and process these grades; they will be kept confidential.
2. TA's grade (**5%**). TA will grade the quality of presentations using Appendix B.
3. Instructor's grade (**5%**). Instructor will grade the quality of presentations using Appendix C.

The presentation mark, broken into the average mark from peers, the marks from TA and instructor, and including comments will be send by email to the groups before the final exam.

Deadlines and Delivery

- Filled in and signed team project contracts must be returned to the instructor by email (to lkurgan@vcu.edu) by **October 15 (Thursday), 2020 before 12:30pm**.
- Submission of the progress report deliverables is due on **October 29 (Thursday), 2020 before 12:30pm**. This includes the training dataset and pdf file with the cover page, description, confusion matrix and MCCs.
- Submission of the final reports and the test predictions is due on **November 19 (Thursday), 2020, before 12:30pm**. The report should be delivered as a pdf file accompanied by the file with the predictions for the blind test dataset.
- Presentation videos in the mp4 format must be sent as a link to the instructor at lkurgan@vcu.edu. The instructor will acknowledge receiving the presentations via a reply email, download the mp4 file, and upload it to google drive to share it with the class – the deadline is **November 24 (Tuesday), 2020, before 9:00am**. **This is a sharp deadline – late submissions are not allowed and will receive 0.**
- Each student will complete a short evaluation form online (Appendix A) to assess presentations of the other groups. This is **mandatory**. You are not allowed to grade your own presentation. Students will have the time **between noon on November 24 and noon on November 25** to mark the videos. The lecture time on November 24 will be designated as the default time to evaluate videos, i.e., the lecture will be cancelled to make room for this activity. The peer evaluation will be based on the average of the marks submitted by the students for a given project team.
- The contribution statements (one per team) must be submitted electronically via email to the instructor at lkurgan@vcu.edu – the deadline is **November 25 (Wednesday), 2020, before 12:30pm**.

Final Notes

- While we recommend the use of RapidMiner, you can complete this project using any other data science software or language. Just make sure that you will be ready to make predictions on the blind test dataset using your selected software.
- 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.
- Your team may be asked to demonstrate how the prediction works, in case of the reported results are irregular. Thus, make sure to retain your software at least until the time of the final exam.
- Always copy the email communications to yourself so you can prove that it was sent.
- Contact the instructor immediately if problems occur.

Appendix A

CMSC 435 Intro to Data Science
Fall 2020
Peer Evaluation Form for the Video Project Presentations

Name of the presenting group ………………………….…………………………………………………

Remarks:

- For each question enter grade between 0 and **30** or between 0 and **10** (0 being the worst, 30 or 10 being the best)
- Optionally please add comments (both positive and negative); they will be passed along to the presenting group.
- Average of these grades submitted for a given group will constitute the peer evaluation component for the project presentation.

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 **30** ……
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 2020
TA's Evaluation Form for the Video Project Presentations

Name of the presenting group ……………………….…………………………………………

TASK	***grade***	*max grade*
Quality of *Motivation for the proposed design*		7
Quality of the *Description of the proposed design*		7
Quality of the *Discussion and comparison of the quality*		7
Quality of *Conclusions*		13
Quality of the Presentation and Presentation Style		16
TA's total mark		**50**

Appendix C

CMSC 435 Intro to Data Science
Fall 2020
Instructor's Evaluation Form for the Video Project Presentations

Name of the presenting group ……………………….……………………………………………

TASK	*comments*	*grade*	*max grade*
Presentation finished within 6 minutes limit	(up to -10 points penalty)		Y / 0
Quality of *Motivation for the proposed design*			7
Quality of the *Description of the proposed design*			7
Quality of the *Discussion and comparison of the quality*			7
Quality of *Conclusions*			13
Quality of the Presentation and Presentation Style			16
	Instructor's total mark		**50**