COMP-551: Applied Machine Learning
Mini-project #1: Predicting the winner!
Due on January 26 2017, 11:59pm.
(Lead TA: Pierre Thodoroff)

Background:

You have been tasked to build a machine learning system to predicting runners' performance in the upcoming 2017 Miami Marathon. We have prepared a dataset, curated from the web, containing past participants' performance in this marathon.

This dataset includes participants in the Miami marathon for the last 15 years.
In the file, the first column contains a unique id for each athlete.
Each rows in the file represents a record of an athlete's performance at a specific race. An athlete can have several records in the database corresponding to different races. The *rank* variable corresponds to the placement achieved by the athlete overall in the corresponding year's competition.
Note that the dataset provided was scraped by us and has not been manually sanitized. It is possible that errors are present (e.g. inconsistent format for the "Time", missing data, etc.).

Your task is to predict two output variables for each participant listed:
- Y1: Will they participate in the 2017 Miami marathon?
- Y2: Assuming they participate, what will be their finishing time? (Submit this even for participates where you predict Y1 = 0).

The 2017 Miami marathon will take place on January 29 2017 (http://www.themiamimarathon.com/). We will compare your predictions to the actual race results!

General instructions:

Using the dataset presented above, use linear methods to predict Y1 and Y2.
Implement and report prediction results for each of the following methods:

a. Predict Y1 using logistic regression, optimized with gradient descent.
b. Predict Y1 using a Naïve Bayes classifier.
c. Predict Y2 using linear regression, optimized in closed-form or with gradient descent (with or without regularization).

You do not need to use the raw variables provided in the dataset, you can encode them as you want, exclude some, supplement with additional features (e.g. weather, race features, etc.), whether scraped from the web or manually added. The data provided in *Project1_data.csv* has been scraped from *https://www.athlinks.com*.

Team organization:

The project must be completed in a group of (exactly) 3 students. You will be required to work with different team members on each mini-project. Please plan accordingly: If you want to do the final project with your best friend, don't work with them for this first project! You can use the class discussion board on *myCourses* to find team members. Anyone auditing the course is welcome to participate in the submission and/or review process. However you should not work with people who are taking the course for credit, to avoid mis-matched expectations.

Submission requirements (1 submission per team, not per individual):

- You must **submit the code** developed during the project. The code can be in a language of your choice. The code must be well-documented. The code should include a README file containing instructions on how to run the code. Submit the code as an attachment (see below).
- You must **submit a prediction file** containing the results of your predictions for Y1 (for both methods) and for Y2 (for linear regression). This file should be in CSV format, 1 line per participant, including 4 columns in the following order: PARTICIPANT_ID, Y1_LOGISTIC, Y1_NAIVEBAYES, Y2_REGRESSION. The only acceptable values for Y1 are 0 and 1 (where 0 = "*no*", 1 = "*yes*"). The values for Y2 should be in hh:mm:ss format. Do not include a header line at the top of your prediction file. Keep the same participant order as in the dataset, starting with participant 0.
- You must **submit a written report** describing your methodology and results. The report should respect the following structure:
    - Project title. (Do not include a cover page.)
    - List of team members, including their full name, email and student number.
    - Introduction: 1-2 sentences describing the problem.
    - Problem representation: Describe what features you use (in a list or table), how you represent them (e.g. categorical, continuous), any design choices (e.g. remove/change some values). If you use additional data, describe what it contains, how it was acquired, how it is represented; include in your report a URL where your dataset can be accessed (do not upload the data on the submission website).
    - Training method: For each of the 3 methods (logistic regression, naïve bayes, linear regression), describe any particular design choices taken when implementing the method. You do not need to include details that are in the class notes (e.g. logistic regression decision criteria, gradient descent algorithm, etc.), unless necessary to understand other details. Include any decisions about training/validation split, distribution choice for naïve bayes, regularization strategy, any optimization tricks, setting hyper-parameters, etc.
    - Results: Analyze validation set results. Describe your evaluation procedure. Present both training set and validation set error. Whenever possible, use figures, tables and graphs to illustrate your work.
    - Discussion: Summarize your results. Speculate about your performance on the unseen 2017 data (test set). Discuss any limitations of your approach, implementation or analysis. Include directions for future improvement.
    - Statement of Contributions. Briefly describe the contributions of each team member towards each of the components of the project (e.g. defining the problem, developing the methodology, coding the solution, performing the data analysis, writing the report, etc.) At the end of the Statement of Contributions, add the following statement: "We hereby state that all the work presented in this report is that of the authors." Make sure this statement is truthful!
    - References (optional). Use appropriate referencing style throughout the report, with the list of references given at the end of the report. References are optional, but should appear if you used any additional data, or adopt methods for feature encoding that were not presented in class, etc.
    - Appendix (optional). Here you can include additional results, longer details of the methods, etc.

  The main text of the report should not exceed 5 pages. References and appendix can be in excess of the 5 pages. The format should be double-column, 10pt font, min. 1" margins. You can use the standard IEEE conference format, e.g. *ewh.ieee.org/soc/dei/ceidp/docs/CEIDPFormat.doc*.

Evaluation criteria:

Marks will be attributed based on: 20% for performance on the 2017 results based on the submitted prediction file and 70% for the written report. Both these components will be assessed per team, not per individual (including late penalties). The remaining 10% is attributed following participation in the peer-review process (i.e. for assessing reports of other groups). This is assessed individually. The code will not be marked, but may be used to validate the other components.

For the prediction results, the performance grade will be calculated as follows: 50% for the Y1 prediction based on accuracy (best of your 2 methods), 50% for the Y2 prediction based on squared-error in terms of Time. For each of these, the team with the best prediction on the 2017 results will receive 100%. Predicting at the level of chance will score 0%. All other grades will be calculated using interpolation between those two extremes (calculated independently for the two output variables, Y1 and Y2).

For the written report, the evaluation criteria include:

- Technical soundness of proposed methodology (feature selection, algorithms, optimization, validation plan)
- Clarity of methodology description, plots and figures (don't forget captions, axes labels, etc.)
- Overall organization and writing (don't forget to spell-check!).

For the peer-review, the instructions and evaluation criteria will be given in class (and included in slides) later; this is not due on January 26.

Submission instructions:

We will be using an online conference management system to coordinate submission of project files and peer-reviews: https://cmt3.research.microsoft.com/COMP551 (The site should be open in a few days.)
You should create an account (one per person) on this site before January 26. You will use your account both as an "author" and as "PC members" (=peer reviewer) throughout the course. Make sure to use the same account for all your activities and submissions. YOU MUST USE YOUR MCGILL EMAIL TO CREATE YOUR ACCOUNT.

When submitting your report, select "Create new submission". Create one submission per group (any of the authors can do this), and link your team members as co-authors. The written report should be submitted as the "Submission file". Only acceptable file format for the report is *.pdf*.

Skip the page on "Edit Conflicts of Interest".

Once your submission is created, from the main console, you will be able to add Supplementary Material. Add the code and prediction files here. The prediction file should be named *predictions.csv*. The code should be aggregated in a single file named *project_code.zip*. Other acceptable file formats for this are *.gz, .tar, .tgz*. Make sure that the code is set up so that we can run it (e.g. include a README file).

You can revise your submission and supplementary material anytime up to the deadline; do not create more than one submission. CMT uses pacific time as the default, therefore the deadline has been set to 9pm pacific time = midnight eastern time. Once the deadline expires, you will not be able to submit files. If you are submitting the project late (subject to automatic 30% penalty), send all files by email to the course instructor.

Final remarks:

As specified in the syllabus, you are expected to display **initiative, creativity, scientific rigour, critical thinking, and good communication skills**. You don't need to restrict yourself to the requirements listed above – feel free to go beyond, and explore further. It is not expected that all team members will contribute equally to all components. However every team member should make integral contributions to the project.

You can discuss methods and technical issues with members of others teams, but you cannot share any code or data with other teams. Any team found to cheat (e.g. use external information, use resources without proper references) on either the code, predictions or written report will receive a score of 0 for all components of the project.