Kaggle Challenge: How many shares?

MA2823: Introduction to Machine Learning

2017

1 Challenge Presentation

Kaggle[1] is a platform dedicated to data mining. In particular, it hosts public data science challenges; the problems are posted by sponsors, and people from all over the world can compete for putting together the best solution.

Your project is centered on the *How many shares* challenge at Kaggle In Class. You can sign up for the challenge using the following invitation URL:

`https://www.kaggle.com/t/0ea3fc897c9748a4b247c89ea50574c8` .

The goal of the challenge is to predict how many times an online article is going to be shared, using variables that describe the article (keywords, topic, length, sentiment analysis, etc.). A detailed description of all features is given on the competition webpage.

For this course, you will enter the competition in teams of 2 to 5 students. Your goal will be to try out algorithms we will see during the course. You are also encouraged to try out other algorithms, combinations of those, and generally explore the challenge data.

The project will be graded on a 2-page report (figures and tables can be moved to an appendix) as well as the ranking of your final submissions to the leaderboard. Final submissions as well as the report are due on **on December 23, 2017, 23:59**.

You are strongly encouraged to use Jupyter + scikit-learn to complete the project, but you will not be graded on your code.

2 Instructions

1. Form teams of 2 to 5 students. **Students who complete their project alone will be penalized unless they have cleared it with me in person.**
2. Register on Kaggle and create a team.
3. Download the data.
4. Set up a cross-validation framework for your analysis of your data.

[1] `https://www.kaggle.com/about`

5. Use this framework for model selection. You are encourage to explore
   - Various machine learning algorithms and their hyperparameters, particularly among those seen in class: (regularized) linear regression, nearest neigbors, tree-based approaches, support vector regression…
   - Various transformations of the features, as suggested in Lab 3 or as you may see fit.
6. Submit as many of those models as you wish (in a limit of 5/day) to the leaderboard to see their performance *on the public validation set*.
7. Submit the predictions made by **2 optimized final models** to the leaderboard. Those are the models you believe should win the challenge. They will be judged according to their performance *on the private validation set*.

3 Evaluation

Final report The report is to be deposited at `http://tinyurl.com/ma2823-2017-hw/` no later than **December 23, 2016 at 23:59**.

Please name your report file "Project_<LastName1><Initial>_<Lastname2><Initial>_<Lastname3><Initial>.pdf" (supposing there are 3 people in your team). If Joe Boyd, Benoît Playe and Mihir Sahasrabudhe form a team, their report should be named "Project_BoydJ_PlayeB_SahasrabudheM.pdf"

Your report should be **no more than 2 pages long.** Figures and tables can be moved to an appendix.

Your report should contain the following elements:

- **Your full names, your Kaggle user names, and your Kaggle team name**.
- A discussion of feature processing. Did you standardize the data, chose alternative representations for some features, discarded other features, and why?
- The cross-validated performance, on the training data, of the models you explored. You are strongly encouraged to explore the space of parameters for each of the algorithms you test. Briefly explain how you do it. Discuss which algorithms/parameters work best.
- The performance, on the validation data (visible part of the leaderboard), of the models you have chosen to submit. Discuss whether the results match your expectations.
- A discussion of additional models you have tried, insights you have gained (e.g. "This method works well but is difficult to fit" or "This method is not very accurate but is really fast to train").
- A discussion of your choice of final model(s). You can submit up to 2 final models. What are these models, how did you construct them, why do you expect them to be your best proposals?

Include tables or figures as you see fit.

Leaderboard ranking The ranking of your team on the leaderboard on **December 23, 2017 at 23:59** will count towards part of your grade. You will find a grading rubric below.

4 Grading rubric

Discussion of feature processing ..... **4 pts**
- Technical quality ..... 2 pts
  Poor / partially misguided / appropriate
- Completeness ..... 2 pts
  Poor / good ideas but could have done more / complete

Discussion of cross-validated performance on the training data **8pts**
- The cross-validation has been properly set up ..... 1 pt
  No / yes
- Multiple algorithms have been tested ..... 3 pts
  0 / 1 / 2 / 3+
- Appropriate hyperparameters have been evaluated ..... 2 pts
  No test / partially appropriate / yes
- Cross-validation results are properly presented ..... 2 pts
  No / partially / yes

Discussion of leaderboard performance **4pts**
- Discussion is technically correct ..... 2 pts
  None / partial or incorrect / complete
- Disucssion is complete ..... 2 pts
  None / partial or incorrect / complete

Discussion of final model ..... **4 pts**
- Discussion is technically correct ..... 2 pts
  None / partial or incorrect / complete
- Disucssion is complete (all obvious points have been addressed) . 2 pts
  None / partial or incorrect / complete

Clarity ..... **5 pts**
- Text ..... 2 pts
  Very unclear / unclear in parts / clear
- Figures ..... 2 pts
  Hard to read / somewhat unclear / clear
- Tables ..... 1 pt
  Hard to read / clear

Final position in the leaderboard **5pts**

TOTAL ..... **30 pts**