ECS607/766 Data Mining Mini-project 2016/17

1 Background

For the course assignment, you will compete in a global online data-mining competition. Kaggle.com is an online aggregator for data mining competitions. The competitions on Kaggle range from classic benchmark problems to cutting edge industrial and scientific problems with real prize money to be won. Kaggle publishes global leaderboards of how well people around the world have done at solving these problems.

This assignment will focus on the easiest problem: Predicting which passengers survived the Titanic disaster (a toy classification problem)[1], and a more difficult bonus problem: Detecting the location of keypoints on face images[2].

You should form teams of 3 students to compete. You must compete in the Titanic competition, and you can optionally compete in the facial keypoints detection competition. *(Note: If you can't find a full team, you can compete in 1s or 2s, but you will not get extra credit for this disadvantage. And you* ***must still*** *create a 1 or 2 person team on the Data Mining QMplus page*[3] *in order to get credit. No team=No mark).*

2 Specifics

You should:

- Form a team of 3. If you are a weak student get a strong partner!
  - It will help if at least one team member knows how to program and/or is an advanced Weka user. You can pass using Excel only, but it will be hard to get high marks.
  - Bachelor and Masters students may **not** be on the same team.
- At least one person from the team should register on Kaggle.
- Read up on both competitions. Download the data for one or both.
- Come up with a solution to the problem.
- Upload your results and see your place on the global leaderboard. You can try different solutions, and upload more times. Kaggle will keep your highest score.
- Do the Titanic competition first. Do the facial keypoints detection competition also for bonus marks.

[1] https://www.kaggle.com/c/titanic

[2] https://www.kaggle.com/c/facial-keypoints-detection

[3] One member of your team needs to register the team on the Data Mining QMplus page. To do this: go to the "Mini-project teams" glossary under the Coursework section, add your team name as an entry and set the list of all team members' student IDs as description.

- Write a report.

3 Structure of your report

Your team needs to write a short (up to half a page of A4 per competition entered, so 1 page max) report on your entry briefly summarizing with bullet points:

- What method you used in your best submission(s).
- If relevant, how you dealt with: under/over-fitting, feature-selection, missing data, outliers, etc?
- What were the hard parts and how you got around them?
- If you entered the facial keypoints detection competition, how did you deal with image data?
- Metadata:
  - Your Kaggle submission name (so we can find your ranking!)
  - Name and student ID of everyone on your 1/2/3-person team.
  - How long you spent working on the project.
  - What each person contributed *(e.g., person 1: scripted, person 2: researched models, person 3: formatted the data)*.
  - Upload your report along with the code you wrote[4] to the Data Mining QMplus page by the deadline (09/12/2016). (Only one team member needs to upload)

4 Grading

- Total: 15% of your final course grade.
- Broken down out of 100 as:
  - Base Mark: 60%. **Criteria**: Successfully submit, and get a better score than "Gender Based Model" baseline on the Titanic competition.
  - Report: 10%. **Criteria**: Documenting everything in report instructions above.
  - Competitive Marks: 30%. **Criteria**: Remaining 30% marks allocated according to your rank[5]. So top rank team within the class gets full 30%, last rank gets 0%.
- Bonus Competition Explanation:
  - Attempting also the bonus (Facial keypoints detection) competition will give you the opportunity to boost your grade.
  - **Criteria**: The class entries for the Bonus competition will also be ranked and marked in the same way as for the basic competition. However, if you enter both competitions, you will be awarded the better of your two ranks/marks for the competitive 30% of the score.

[4] Save the report as a pdf file, include the source code (no binaries) or spreadsheets you used and compress everything into a zip file that you will upload. Use the name of your team as the name of the zip file.

[5] We will download the current Kaggle scoreboard on the submission deadline.

  - **Criteria**: Successfully submitting, and getting a better score than the baseline on Facial keypoints detection that predicts each keypoint location as the mean of that keypoint's locations from the training set[6]: Add 10% bonus to all the above (up to the 100% cap overall).

Notes: This grading scheme means that:

- Everyone should be able to get 70%, as this is the lowest mark (last place) for minimum project criteria (submit & document a reasonable solution to the Titanic competition).
- Everyone can get 80% (even if last place), just for submitting both tasks (70%+10% bonus).
- You can still potentially get 100%, even if last on one competition (by being first on the other competition).

5 Notes

- You have a couple of weeks to complete the project. It is doable, but **do not wait until the last couple days to get started**. People who do this struggle and may fail.
- Kaggle is keeping the true test data hidden. So a perfect score on the data you download doesn't necessarily mean a great score when you upload! (Overfitting!)
- There are lots of forum boards on Kaggle where people discuss techniques and tips and tricks. Read them to get advice and real world tips and tricks.
- You are free to use whatever software you want for the data mining: Excel, Matlab, Weka, Python, R, etc. Matlab has KNN/DecisionTrees/SVMs/etc built in. Python with machine learning extension (scikit-learn[7]) has even more built in.
- Solutions may involve lots or no programming depending on your preference and how you approach the problem.
- You may need to convert data formats: (i) from what you get from Kaggle to a format readable by your favorite programming language / data-mining software and (ii) from the output of your mining back to the results format required by Kaggle.
  - This may not be a trivial depending on how you approach the problem. **Solving this is part of your task and learning experience**. The scripting and/or excel hacking required for this are the "ugly" side of practical data mining. But this is exactly what you will have to deal if you do real-world data mining in industry!
- This project is worth trying hard on. If you have a good Kaggle score, and apply to work in a data analytics related job, your Kaggle rank is something you can put on your CV!
- Good luck! :)

[6] The score achieved by this approach will be 3.96244. An implementation of this approach in R is available at: https://www.kaggle.com/c/facial-keypoints-detection/details/getting-started-with-r. Note that submitting the other solution from that page that gives a score of 3.80685 will not count as your own.

[7] http://scikit-learn.org/stable/