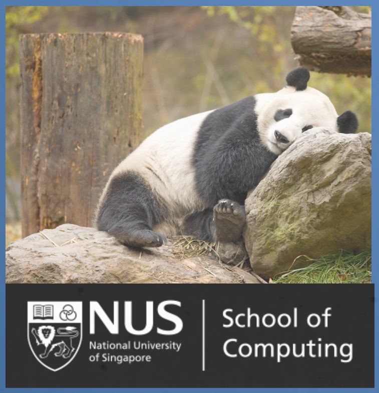

Machine Learning

NUS SoC, 2017/2018, Semester I
Lecture Theatre 19 (LT 19) / Tuesdays 14:00-16:00

Last Updated: October 10, 2017 - Initial page port.

Homework #3 » Kaggle Competition

In this homework assignment, you will be participating in a competition to create the best learner that can predict a target function. In contrast to the previous two individual assignments, you are encouraged to officially collaborate with up to two (2) other peer classmates to form a team of maximum size 3 to work on this assignment.

You have a choice of two competitions, both which are detailed and hosted on Kaggle.com. You will need at least one team member to sign up for a Kaggle account to work on the homework assignment. If you are unwilling to create an account on Kaggle (for any reason, such as privacy) please contact the CS 3244 staff so that we can find an alternative access pathway for you:

- Rossmann Store Sales (https://www.kaggle.com/c/cs3244-rossmann-store-sales). This is a time series prediction regression task.
- Labeled Faces in the Wild (https://www.kaggle.com/c/labeled-faces-in-the-wild-2017). This is an image recognition machine learning task.

Access to the datasets and important details on each task are available at the respective competition's websites. Details will be only updated on the respective webpages on Kaggle.

Two important details about the deliverables:

1. README.pdf. Your README (in .pdf) needs to contain a short 1-3 page report detailing how your team tackled the task, inclusive of documenting any data analysis you performed on the data, normalization, models tried, and any post analysis that you did. In addition to the report, you need to also describe the files that are submitted, include your statement of independent work (or suggest alternative evaluation criteria), work allocation among team members (applicable to teams with 2 or 3 people), and a section

of the references you used in your project. Please also put the name of your team in the beginning of the report, so that we can associate your team with your leaderboard position on Kaggle.

2. Held out testing set. Kaggle makes both the training and testing data available for you to download. Since both problems' full datasets are publicly available, you could recover the labels for the test sets that we provide you, but this would be cheating. This is not allowed and all teams should not attempt to use the original public datasets to reconstruct the testing data that is hosted on Kaggle. Such cheating defeats the purpose for the assignment, and your team would be better off investing effort in learning how to apply ML appropriately instead of wasting time to circumvent our efforts to give you a genuine learning problem. Offending teams will be severely prosecuted.

   We will also construct new data on our own which we do not provision to Kaggle (held-out testing data), which is given an equal weight in the grading.

   We will be asking you to demonstrate your team's learner on new instances within X days after Homework #3's submission deadline, to establish held-out testing performance. These dates and the workflow process details will be announced later.

Submission Formatting

For us to grade this assignment in a timely manner, we need you to adhere strictly to the following submission guidelines. Following the submission guideliness will help us grade the assignment in an appropriate manner. You will be penalized if you do not follow these instructions. Your matric number in all of the following statements should not have any spaces and all letters should be in CAPITALS (inclusive of the ending check letter). You are to turn in the following files:

- Your source code files `hw3-(rossman|lfw).(py|ipynb)` for this homework assignment. We will be reading your code, so please do us a favor and format it nicely. To expedite grading, we need to require all of the assignments to use Python 3.4 (3.5 should ok as well). Do not put your high level documentation in this file, but place it in the `README.pdf` file instead.

- A .PDF file `README.pdf` that contains your README, complete with 1) report with high-level description, 2) file listing, 3) statement of teams' independent work, 4) references.
- Any ancillary files that your submission uses. We may or may not read them, so please put the core part of your programming answers in the `hw3-(rossman|lfw).(py|ipynb)` file.

Please do not submit the training and testing files, as this will significantly blow up your submission's size and cause unnecessary bandwidth consumption.

These files will need to be suitably zipped in a single file called `<matric number>.zip`, `<matric number>` is replaced by a string of team members' matric numbers strung together, joined by dashes ('-'; eg. `A0000001X-A0000002Y-A0000003Z`), sorted lexicographically. Please use a zip archive and not tar.gz, bzip, rar or cab files. Make sure when the archive unzips that all of the necessary files are found in a directory called `<matric number>` (note to Mac users please test your zip archives on the command line; as OSX "automagically" creates the parent folder but it is often actually not contained in the archive). Upload the resulting zip file to the IVLE workbin by the due date: 7 Nov 2017, 11:59:59 pm SGT. There absolutely will be no extensions to the deadline of this assignment. Read the late policy if you're not sure about grade penalties for lateness.

Grading Guidelines

Grading criteria is common to both assignments and given on the respective competition's site. We replicate it here for your convenience:

Documentation, which consists of: • 20% on the documentation itself: formatting, clearness, comprehensiveness of your README, as well as how well your team complies with the competition rules. • 20% on the creativity, insight or depth of your solution.	40%
Relative Performance: • 10% on Training data performance (as reported by you) • 25% on Given testing performance (as reported by Kaggle) • 25% on Held-out testing performance (not available for testing against in Kaggle, will be compiled by our TA group.	60%
Total	**100%**

Hints

- We strongly suggest that you register for a Kaggle account early and attempt to use the framework to submit a trial submission. Iron out the kinks in your submission processes early.

- Kaggle has some interesting parts of its site that deal (much) more with the practical aspect of fielding machine learning. You may find it useful (even fun!) to participate in other competitions and learn how others are working on the problems by using their pre-cooked solutions (kernels) and the competition forums. Jupyter notebooks also play a role in the site, so do spend some time getting familiar with how other professionals approach learning from data, in order to come up with your strategy.
- Both of our homework #3 variants are byproducts of other (larger scale) machine learning competitions. You can look up the relevant papers to explore how other scientists approach these problems.
- Both assignments are graded independently of each other. You are only competing against other class teams working on the same assignment. In the case where it becomes apparent that one assignment is markedly harder than another, we reserve the right to balance the grading to reflect this difference in whichever way we choose.

Back to top