Peer grading assessment of machine learning exercises in Aalto University

Shaghayegh Safar, Alexander Jung
Department of Computer Science, Aalto University, Finland
firstname.lastname@aalto.fi

Abstract

The grading of exercises is a core activity for the evaluation of students performance in many academic courses. During recent years. machine learning related courses at Aalto University, including the course CS-E3210 *Machine Learning: Basic Principles (ML:BP),* have attracted an increasing number of students which peaked at more than 800 enrollments for ML:BP in 2018. This increase in student numbers makes the manual evaluation (by teaching assistants) of exercises assignments difficult. An alternative to manual evaluation (grading) of exercise assignments is provided by peer grading software platforms. These platforms allow to organize peer grading, i.e., students evaluate the solutions of their colleagues. In this paper we report some insights gained while implementing peer grading in ML:BP. In particular, we analyze students feedback collected during and after the course. We further report a summary of recommendations for using the peer grading system in the future based on the students feedback and interviews with six teacher assistants that worked with the system.

Introduction

Peer grading[1] is a system of grading where students have the main role in the assessments. The procedure can be summarized as: 1) the course instructor prepares the assignment and the rubric questions (i.e., a set of questions that the grader will answer about the submitted solution). 2) The student submit their anonymized solutions for the assignment. 3) Each submission is graded by several students through answering the rubric questions. 4) The students observe the given feedback and engage in discussion and flag the assessments that they think are unfair. 5) The course instructor will go through the flagged assessments and makes the final decision about the grade.

The system has been used in Aalto University in several machine learning related courses. It has been used in one semester in Machine learning: Advanced Probabilistic Methods (ML:APM) and Machine Learning: Basic principles (ML:BP), and more than one semester in Bayesian Data analysis (BDA), during the years 2015 - 2018. On average, it seems that the outcome was not at a satisfactory level for ML:APM and ML:BP; However, the system has been used in BDA for several semesters. In this report we investigate the peer grading system and report its

[1] The report is based on the experience with the peergrade.io service (https://www.peergrade.io/).

advantages and disadvantages. This report may be helpful for people who are interested to use this system in machine learning related courses.

Problem Setup

The number of students who are interested in taking machine learning related courses have increased substantially in Aalto University in recent years. For example, the course ML:BP, which is an introductory course to the field, has gained a substantial interest through the years as depicted by figure 1.

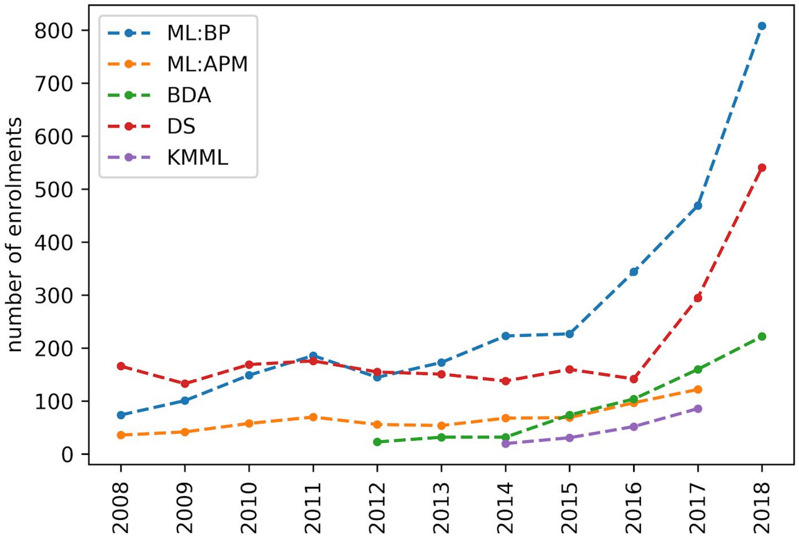

Figure 1. Number of enrolled students in machine learning related courses in Aalto University between 2008-2018. There is a significant increase in basic machine related courses such as ML:BP and Data Science (DS) in recent years.

This increase has caused an unexpected need in number of teacher assistants to help the teacher to organize the exercises, and more importantly, for grading of the students submissions. For ML:BP, on average there are 4-6 round of exercises and one final project that needs to be submitted by the students throughout the teaching period. Furthermore, the data indicate that even for more advanced level courses such as Kernel Methods in Machine Learning (KMML), ML:APM, and BDA, the number of students have multiplied during the last four years (increasing from 20 to 86 for KMML, 68 to 122 for ML:APM, and 74 to 222 for BDA). Peer gradings seems to be a good candidate solution to tackle the increase in number of students and organization challenge of these courses.

Previous studies have shown that students who engaged in peer-grading performed better on subsequent tests than did students who did not [1]. Furthermore, the peer-reviewed scores can be equivalent to expert scoring [2].

In this report, rather than testing or confirming these hypotheses, we investigated how the peer grading system has been implemented in practice in machine learning courses and then we recommend points to further improve the implementation in the future. This report is built on the evaluation of 300 feedback from students who had the ML:BP course using peer grading and interviewing six teacher assistants (TAs) involved in the mentioned courses. We tried to find what were the main reasons of failure and success in the previous courses. The results are reported as a list of questions and answers. We summarize the important points for preparing a guideline for grading. It should be noted that we did not perform a controlled study and did not measure the added gain by the system compared to not having the system, but rather we just summarize a set of suggestions that can be helpful in the implementation of peer grading for future courses.

Questions and answers

Does the peer grading help the students to understand the subject better? The main reason for using this system was learning the subject better by grading others' assignments and seeing different types of answers. From this aspect, both the majority of students and TA's agreed that the system was successful. Using the system, the students can learn to think differently and from a new perspective and to see how they can reach the answer by using different approaches.

Is peer grading accurate? Each assignment needs to be graded by more than one person. Also if there is any problem and unfair assessment, the students can use flags to ask course assistants to look into it. Therefore, peer grading can be as accurate as instructor grading. The only potential bug in this system can happen in the case where students do not grade assignments critically. For example, if all students agree among themselves to give positive feedback to each other then there would be no conflicts and all students would be happy.

Should there be rewards for good feedback and penalties for bad ones? From the student's point of view, some feedbacks are very short, not relevant, wrong, and useless. These feedbacks will be flagged and the TA's should look into them and do the grading from the beginning. Also, some students flag every question which they received a bad grade. This behaviour will add more work for TA's. To avoid this, there should be a serious penalty for those who give a wrong grade and for those who give a wrong flag.
In the same sense, rewards have a good effect too. Bonus points encourage students to grade correctly and precisely. However, it seems that penalties are more effective than the rewards.

How peergrading affects the workload of students and instructors? In the students feedback, one of the biggest complains was about the workload. They said that they spend too much time for just grading and this added workload caused problems for them since they usually had more than one course. It seems that having a detailed and clear guideline (rubric in the peergrading system) may help to reduce the workload.
One of the advantages of the peer grading system is reducing the workload of instructors. However, if there are many flags this cannot happen as a TA may need to re-evaluate a question of a submission more than once in case that question was flagged several times. Again, having an accurate and explicit guideline will decrease the trouble for TA's.

How many peer graders are needed per assignment? 3-4 persons.

How many assignments each student should grade? Based on the TA's opinion, the ideal number of assignments for each student is 3-4.

How to prepare a good guideline:

Rubrics (guidelines) are a list of questions that the grader answers based on the submitted assignment. For example, "has the submission found the value x=3?". The rubric is the most important part in peer grading system. Based on TAs interviews, the following points should be considered while preparing the rubric questions:

1. **Designing the assignments:**
    - While designing the assignments, have in mind that the assignment will be graded based on a rubric guideline and not by a TA.
    - Design the assignment parts such that they are independent from each other. For example, when the output of question 1 is necessary for question 2, it would be extremely hard to design rubrics since the solution of question 2 cannot be graded only objectively.
    - In the assignment, mention explicitly what the student should provide as output and what exactly will be evaluated in grading. For example, sometimes the student may just write the code or provide the final figure and think that it is enough. However, the rubric may evaluate a number from the figure that the student has not written explicitly in the submission.
    - Do not give too much flexibility to students for their answers. For example, don't use something like "repeat the process for different values of the hyperparameter and report the best result", instead, please mention explicitly some values that should be tried out.

2. **Designing the rubric for assignments**:
    - The rubric questions need to be very precise/explicit/unambiguous, otherwise there will usually be many flags (if the question or grading can be understood in many ways).
    - In multiple choice questions one has to be careful that one of the choices always holds. For instance, consider a question:
      "Is there a graph that includes both a cat and a dog?" a) Only cat plotted b) Only dog plotted c) Both cat and dog plotted. In this case the alternatives are incomplete since there would also need to be the choice "d) neither is plotted". This might seem a too simple example but in practice TAs have observed that it is very easy to design a question that can be ambiguous or that there is no clearly correct choice for the evaluator to pick from.
    - Have in mind that students can easily overfit to the final grading system and in some cases they may skip some of the middle steps of the assignment. It would be best if a question can naturally be divided into several steps for grading.

3. **Communication with the students:**
    - Everything has to be explained clearly and thoroughly to students. This increases the initial workload, but will significantly reduce the TA work afterwards. Be consistent in whatever you decide to do and communicate clearly to students exactly what is expected of them in every single part of peergrading. This concept is new for many students and many students did not fully grasp of their duties in peergrading. Ideally, consider one session to go through all the steps with students and explain clearly the rewards and punishments. Let the students know that peergrading is time consuming so that they would build the right expectations.
    - without some serious motivation (bonus points and serious penalties) there will be plenty of students who will not spend more than the absolute minimum of time needed, resulting in very noisy grades. and lots of flags afterwards. These also need to be communicated clearly to students.

Acknowledgements

We would like to thank the teacher assistants of the mentioned courses who provided valuable feedback about their experience with the peergrading system. Furthermore, we would like to thank prof. Aki Vehtari for providing comments and data about the peergrading experience.

References

1. Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, *109*(8), 1049.

2. Price, E., Goldberg, F., Robinson, S., & McKean, M. (2016). Validity of peer grading using Calibrated Peer Review in a guided-inquiry, conceptual physics course. *Physical Review Physics Education Research*, *12*(2), 020145.