LH216V assignment 2: Final grading criteria

Submitted by: Gustav Eje Henter

Course code: DD1420
Course name: Foundations of Machine Learning
Number of credits: 7.5
Course offerings: Periods 1 and 3 every academic year
Number of students: Approximately 50 MSc students in period 1, approximately 30 BSc students in period 3. These numbers are expected to grow.

Note: These are not the actual grading criteria of course DD1420, but merely a proposal.

Part 1: Part of a course memo

1.1 Intended learning outcomes

After passing the course, the student should be able to:

- *interpret basic concepts, language, and notation that supports machine learning*
- *use mathematical and statistical methods that underpin machine learning*
- *derive and prove selected theoretical results*
- *implement machine-learning models to solve empirical problems*
- *interpret the results of applying machine-learning models to data*
- *exercise critical thinking around ethical, societal, and sustainability aspects of machine learning and reflect on how one's professional activities might influence these aspects now and in the future*

in order to

- *be able to define problems in data analysis clearly*
- *formulate a suitable solution with machine learning and strengthen this solution through critical and quantitative evaluation*
- *be well prepared to take advanced courses in machine learning.*

1.2 Grading criteria

Please turn to the end pages of this submission, where the grading criteria are specified in Table 1. Associated with that table are the following notes:
1) For cases where a single criterion maps onto multiple letter grades, fulfilling the criterion counts as the higher grade.
2) It is implied that interpretations, calculations, derivations, implementations, interpretations, etc. must be correct to fulfil each criterion.

The same criteria and policies apply to BSc and MSc students equally, across the entire course, always.

1.3 Assessment tasks

The course is assessed through:

- Individual proctored summative Canvas quizzes (graded automatically). These contain two different kinds of problems:
    - Computation (“formula”) questions with numerical answers
    - Non-numerical questions, e.g., multiple choice
- Assignments completed on Google Colab by students working in pairs. These assignments contain three different kinds of problems:
    - Theory problems (graded by teaching assistants)
    - Coding problems (graded automatically)
    - Interpretation problems (graded by teaching assistants)
- Participation in a seminar or satisfactorily completing a designated back-up activity

Except for multiple choice questions, your score on each problem type (formula, theory, coding, and interpretation) is summed across all modules for a total score in each category. That score is used to assess how well you meet the different learning criteria in the course and, thus, your final grade (F through E).

1.4 Alignment between assessments and grading criteria

This is clarified through operational grading criteria, which are enclosed in table format at the end of this document.

1.5 How results affect the final grade (a formula for computing the final grade)

If you do not achieve a pass on all learning outcomes in the course, a final grade of F or Fx is awarded. Fx is only awarded to students that fail one single learning outcome but achieve a pass on all others. Failing more than learning outcome results in an F.

If you achieve a pass on all learning outcomes in the course, your final letter grade is based on your lowest letter grade on any individual grading criterion. However, if you meet the criteria for a higher grade on at least half of the learning outcomes, a grade one above your lowest letter grade is awarded instead.

Examples of letter grade distributions across five outcomes and resulting final grades:

- 3 A, 2 B: Grade A awarded
- 2 A, 1 B, 2 C: Grade B awarded
- 1 A, 2 B, 2 C: Grade B awarded
- 3 A, 1 B, 1 D: Grade C awarded

Students who receive an Fx will be given an opportunity to take an oral exam to demonstrate their fulfilment of the specific learning outcome that they initially failed. If their performance on that exam suffices for achieving a pass on the outcome in question, they will receive the grade E on the course.

Part 2: Reflections

2.1 Motivating the design of grading criteria and assessments

I will first describe the assessment activities in the course (and their motivation), and then move on to describe my thinking in deciding how student performance is mapped onto letter grades for each learning outcome.

As background, the course is intended as the first course in machine learning specifically for students who want to become experts in that field, and thus covers a significant amount of ground on diverse topics in machine learning. It is especially intended to prepare students for advanced machine-learning courses at KTH. In particular, it is mandatory in year 1, period 1 for the machine-learning MSc programme at KTH (TMAIM).

Since the TMAIM MSc programme sees high application pressure and has many students, the assessments are all designed with easy scaling in mind. For scalable summative assessment, there are automatically graded Canvas quizzes (TES1 in Ladok), which students complete individually, and assignments (INL1 in Ladok) in the form of Google Colab notebooks, which students work on and submit in pairs. For Canvas quizzes, certain questions require students to perform a computation and input a numerical answer; scores on such questions can be tracked separately using the outcomes functionality in Canvas. Coding problems on the assignments are automatically graded using the `nbgrader` tool for Jupyter Notebooks, whilst assignment problems that require students to perform proofs or derivations or written motivations are scored manually by TAs.

The main course materials are hypertext lecture notes on Notion and corresponding video lectures. A large amount of the learning occurs by students working their way through the assignments, and by completing automatically graded formative practice quizzes on Canvas. If students have questions or need help, they can post questions on Canvas message boards for input from teachers and teaching assistants, as well as other students. Alternatively, students can sign up for 20-minute one-on-one help-session slots with teachers or TAs.

The course is divided into 9 modules and uses continuous assessment. Each of the 9 modules comes with a quiz and an assignment, which have to be submitted at the end of the module. This means that there are one or two instances of each type of assessment each week.

Because points are summed across all 9 individual assessment activities to decide each letter grade, one can maintain a high resolution (many different letter-grade results) for most grading criteria. The exception is the first learning outcome, which only considers the number of course modules (out of 9) for which the student achieved a passing grade.

My formula for combining grades (which I reflect on and motivate further down) rewards consistency, across topics/modules but especially across different tasks/skills/learning outcomes. Specifically, the final letter grade is strongly limited by the lowest letter grade on any learning outcome. My philosophy has therefore been not to make the requirements to get a good grade on any individual learning outcome not particularly demanding.

The proposed score thresholds in the operational grading criteria are based on the current score distributions on quizzes and assignments in the course. For example, since we provide test cases for the coding problems, allowing students to verify their solutions before they submit, nearly all students get the coding parts of the assignments correct (as measured by our hidden, automated test cases). Requiring a high degree of correctness to the learning outcome related to coding is thus not overly demanding, not is it overly demanding to require 95% correctness for an A. Conversely, some theoretical derivations in the course can be difficult and may only be solved by strong students. The number of points on theory questions required for a passing grade, in relationship to the maximum possible grade, is therefore quite low. As a final consideration, automatically graded formula questions (with numerical responses) in Canvas cannot disregard simple typos or other slip-ups unrelated to true understanding, and the score for a maximum grade there is therefore more lenient.

In practice, I would like to see the effect of these criteria by running the two systems in parallel for one iteration (as was done at Uppsala University), to tune criteria and thresholds before implementing them in practice. But before any changes can be made, it is essential that the other two lecturers in the course agree to modify the grading system. What is contained in this document is a thus just a proposal at this stage, and nothing more.

2.2 Considering peer feedback

I did not receive a lot of actionable feedback on my submission for assignment 1a. The most concrete feedback pertained to the ILO “After passing the course, the student should be able to interpret the adaptation of the model to data”, where a peer reviewer wrote: “I'm not entirely sure how the verb ‘interpret’ in this context is going to be assessed. Indirectly, in the sense that the student needs to interpret the questions to be able to answer them?” The ILO was therefore changed to “After passing the course, the student should be able to *interpret the results of applying machine-learning models to data*” (italicised here for clarity).

Regarding my submission for the preliminary version (2a) of this assignment, another peer reviewer wrote: “I am wondering if the criteria and tasks are the same for the BSc and the MSc students. It might be useful to clarify that.” A clarification has been added to Section 1.2 above. That reviewer also requested more information on the learning activities for students in the course. A paragraph about this has been added to Section 2.3 below.

One of the course teachers provided several comments on the preliminary version (2a) of this assignment. Importantly, they noted the absence of the required specification for awarding the Fx grade. Information about that, and how students may convert an Fx grade into an E, have now been added to Section 1.5. They also pointed out that Table 2 with operational grading criteria could be made simpler and easier to read by splitting out common formulations. I agree, and this has been implemented in this version of the submission. Doing so also helped remove a minor inconsistency due to previous copy-pasting, so it was clearly a good idea. The grading criteria for the first ILO in Table 1 have also been updated to refer to “course modules” rather than the “course” in abstract, again based on feedback from the same teacher.

2.3 Reflections on using the grading criteria in teaching

To make students aware of what is expected of them at the assessment, I will highlight the grading criteria, including operational criteria, in the course memo, and then spend time introducing them at the first lecture of the course (also recording video so that the information is available to those who miss the lecture). Beyond that, each assessment clearly spells out how many points it is worth across the different question categories, in relation to the total for these problems across the course. Each problem is labelled with its category (formula question, theory question, coding question, interpretation question) and the maximum achievable score. Examples are provided to illustrate the formula used to compute the final score.

For theory questions and interpretation questions, where the score cannot immediately be calculated automatically and instead require manual work by a TA, TAs will use a standardise whitelist/blacklist on what needs to be present for a specific score on each problem. This will help reduce the amount of subjectivity in the scoring process. For the benefit of the students, a few example questions (not used for actual examination) will be provided in the course memo or on Canvas. These will show the question asked, the otherwise-secret whitelist/blacklist used to score the answers, example responses, what scores they were given, and why.

2.4 Reflecting on the effects of the formula for combining grades

Currently, the grade distribution on the quizzes in the course is reasonably spread out, whereas nearly all students achieve high scores on the assignments. Given this, requiring consistent performance for higher grades does not feel unduly burdensome on the students, especially if the demands for each individual criterion are not overly demanding. It also feels appropriate that a good grade in a course on machine-learning fundamentals indeed means that one has a grasp on all fundamentals, without significant blind spots. After all, this course is specifically designed for students who specifically aspire to become professionals in machine learning, and thus need a solid foundation to build upon.

Because of the continuous assessment, it is difficult to reach higher grades without completing all quizzes and assignments, so there is little risk that students feel that their grade is locked in and stop caring near the end of the course. Even if students struggle with a particular learning outcome, meaning that they know that that outcome will limit their final grade, all assessments count in aggregate, so being less ambitious near the end of the course is risky, and the fact that they can score higher than their worst grade if they do better on most other learning outcomes encourages them to keep trying to learn the material and demonstrate good performance on each outcome.

Table 1: Grading criteria in matrix form

<table>
<tr><th>Learning outcome</th><th>E</th><th>D</th><th>C</th><th>B</th><th>A</th></tr>
<tr><td rowspan="2">interpret basic concepts, language, and notation that supports machine learning</td><td>…throughout well over half of the course modules</td><td colspan="2">…throughout nearly all of the course modules</td><td colspan="2">…throughout all course modules</td></tr>
<tr><td colspan="5">Assessed using summative Canvas quizzes</td></tr>
<tr><td rowspan="2">use mathematical and statistical methods that underpin machine learning</td><td>…in at least half of the cases considered in the course</td><td>…in marginally over half of the cases considered in the course</td><td>…in well over half of the cases considered in the course</td><td colspan="2">…in many of the cases considered in the course</td></tr>
<tr><td colspan="5">Assessed using formula questions on summative Canvas quizzes</td></tr>
<tr><td rowspan="2">derive and prove selected theoretical results</td><td>…in a few of the cases considered in the course</td><td>…in close to half of the cases considered in the course</td><td>…in marginally over half of the cases considered in the course</td><td>…in well over half of the cases considered in the course</td><td>…in many of the cases considered in the course</td></tr>
<tr><td colspan="5">Assessed using theory problems on Google Colab assignments</td></tr>
<tr><td rowspan="2">implement machine-learning models to solve empirical problems</td><td>…in well over half of the cases considered in the course</td><td>…in many of the cases considered in the course</td><td>…in close to nearly all of the cases considered in the course</td><td>…in nearly all of the cases considered in the course</td><td>…in virtually all of the cases considered in the course</td></tr>
<tr><td colspan="5">Assessed using coding problems on Google Colab assignments</td></tr>
<tr><td rowspan="2">interpret the results of applying machine-learning models to data</td><td>…in marginally below half of the cases considered in the course</td><td>…in at least half of the cases considered in the course</td><td>…in marginally over half of the cases considered in the course</td><td>…in well over half of the cases considered in the course</td><td>…in many of the cases considered in the course</td></tr>
<tr><td colspan="5">Assessed using interpretation problems on Google Colab assignments</td></tr>
<tr><td rowspan="2">exercise critical thinking (…)</td><td colspan="5">…by actively participating in discussion of these topics at the designated seminar or completing a similar learning activity</td></tr>
<tr><td colspan="5">Pass/fail only; no letter grade. Assessed using seminar or (in case of student absence) a designated backup activity</td></tr>
</table>

Table 2: Operational grading criteria

<table>
<tr><th>Learning outcome</th><th>E</th><th>D</th><th>C</th><th>B</th><th>A</th></tr>
<tr><td>interpret basic concepts, language, and notation that supports machine learning</td><td>≥70% score (pass) on at least 6 of 9 quizzes</td><td colspan="2">≥70% score (pass) on at least 8 of 9 quizzes</td><td colspan="2">≥70% score (pass) on all quizzes</td></tr>
<tr><td rowspan="2">use mathematical and statistical methods that underpin machine learning</td><td>≥50% of maximum possible score…</td><td>≥60% of maximum possible…</td><td>≥70% of maximum possible score…</td><td colspan="2">≥80% of maximum possible score…</td></tr>
<tr><td colspan="5">…summed across the formula questions on all quizzes</td></tr>
<tr><td rowspan="2">derive and prove selected theoretical results</td><td>≥35% of maximum possible score…</td><td>≥45% of maximum possible score…</td><td>≥60% of maximum possible score…</td><td>≥70% of maximum possible score…</td><td>≥80% of maximum possible score…</td></tr>
<tr><td colspan="5">…summed across the theory problems on all assignments</td></tr>
<tr><td rowspan="2">implement machine-learning models to solve empirical problems</td><td>≥70% of maximum possible score…</td><td>≥80% of maximum possible score…</td><td>≥85% of maximum possible score…</td><td>≥90% of maximum possible score…</td><td>≥95% of maximum possible score…</td></tr>
<tr><td colspan="5">…summed across the coding problems on all assignments</td></tr>
<tr><td rowspan="2">interpret the results of applying machine-learning models to data</td><td>≥40% of maximum possible score…</td><td>≥50% of maximum possible score…</td><td>≥60% of maximum possible score…</td><td>≥70% of maximum possible score…</td><td>≥80% of maximum possible score…</td></tr>
<tr><td colspan="5">…summed across the interpretation problems on all assignments</td></tr>
<tr><td>exercise critical thinking (…)</td><td colspan="5">By being present at the designated seminar or submitting a satisfactory back-up assignment</td></tr>
</table>