MOOC Data

Matt West, April 8, 2014

Data overview: This data is the assessment information for students from a recent MOOC (Massively Online Open Course). It includes per-question data from 8 weekly quizzes as well as per-question data from the final exam.

Course information:

- Title: "Introduction to Sustainability"
- MOOC platform: Coursera
- Website: `https://www.coursera.org/course/sustain`
- Instructor: Jonathan Tomkin, University of Illinois at Urbana-Champaign
- Session: Repeat 4, started August 26th, 2013
- Length: 8 weeks

Questions to ask:

1. Can we predict final exam performance?
2. Can we predict who will drop the course and when this will happen?
3. Do quiz scores correlate with the corresponding questions on the final exam?
4. Do scores on quiz questions in each Bloom's taxonomy level correlate with the same levels on the final exam?
5. Can we predict student answers on a per-question basis?

Levels of detail:

1. Overall scores assessment items (1 score per quiz, exam, etc).
2. Per-question correct and incorrect information for each quiz and final exam.
3. Per-question answers for each quiz and final exam.
4. Other data, such as submission times, orientation quiz score, enrollment order, etc.

Data description:

- Each MOOC student has a unique `coursera_user_id` that links student data between different assignments.

`gradebook.csv`: The summary gradebook information.

- Column A: the `coursera_user_id`
- Column D: score on an orientation test about the organizational structure of the course.
- Columns E–T: Quiz scores, 2 quizzes per week. Detailed information on the first quiz in each week is in separate files. No detailed information is available for the second quiz in each week.
- Column U: Final exam score. Detailed information is in a separate file.
- Columns V–X: Discussion scores from weeks 3, 5, and 7 (maximum score: 80).
- Some assessment items can have partial points awarded for late submission. For example, student 54684 has 6.5 points for Week 1 Quiz 1 in `gradebook.csv`, but in `week1quiz1.csv` they got 13 questions correct (half points awarded for a late submission).

`week[1-8]quiz1.csv`: Per-question data for the first quiz in each week.

- Column E: the `coursera_user_id`
- Columns F–end: per-question response data:
  - Column `Q01` is the answer number (1–5) given to question 1 (9999 indicates "not answered").
  - Column `Q01_corr` is 1/0, showing whether the answer to question 1 was correct/incorrect.
  - Some questions have multiple variants and each student is asked a random variant, in which case there are multiple answer columns (`Q01v1`, `Q01v2`, etc), only one of which contains a valid answer and the others of which are marked 9997.
- In each quiz the questions correspond to different levels in Bloom's taxonomy:
  - Questions 1 to 8 are "Knowledge"
  - Questions 9 to 14 are "Comprehension"
  - Questions 15 to 20 are "Application"

`finalexam.csv`: Per-question data for the final exam.

- Column G: the `coursera_user_id`
- Columns H–end: per-question response data in the same format as the quizzes.
- Final exam questions correspond to quiz questions as follows:
  - The 40 final exam questions have 5 questions for each week (so Questions 1–5 are from week 1, 6–10 are from week 2, etc).
  - Within each group of 5 questions, the first 2 are "Knowledge", the second 2 are "Comprehension", and the last is "Application". Thus Q11, Q12 are Knowledge questions for week 3, Q13, Q14 are Comprehension questions for week 3, etc.