Student Career Main Lab

Computer-Aided Simulations Lab G5

I. Problem overview

Our goal with the present task is to enhance the previous implemented simulator for a student's career, where we evaluated the graduation time and final grade at a MSc course, specifically at Politecnico di Torino. The proposed improvements here are:

- Individualize approval probability and grades distribution according to each course, by collecting data from Polito website "Statische superamento esami";
- Allow students to reject a mark according to a stochastic event;
- Allow students to take two appeals of an exam on a specific session once a year.

II. Proposed approach

A. Data sources and data collection

The chosen data source is the Polito website, where each course has a dedicated web page where we can find the approval statistics and grades distribution. To access and extract this information, we use the packages `request` to get the HTML content of the page and `re` to search for RegEx (regular expressions) on the page.

To get the number of people who took and who passed an exam, we search for the expressions `\Totale iscritti:\`and `\Superi:\`, respectively. As for the grades distribution, on the HTML page, there is a list with the count of people who achieved each grade from 15 to 30 on the line after the one where the expression `\name: 'Iscritti',\`is found.

Finally, to evaluate a student career, we start from a list of courses, which can be given as a list or a path to a text file containing one course on each line. The courses may be referred as its code or the URL to the exam statistics website. From this list, we instantiate one object of the class `Course` for each course in the student career.

B. Stochastic events

There are 4 different random elements implemented in the simulation: *(1)* number of exams taken by a student during each session; *(2)* the success or not of a student at each exam; *(3)* the grade a student achieve at a passed exam and *(3)* the rejection or not of a mark received by the student.

1) Exams taken during a session: The chosen approach for the number of exams taken per session was to draw it from a binomial distribution. In this case, consider that the student takes, on average, a given number of exams $E[X]$ and the probability of taking each exam is $p \approx 50\%$. By doing that, we define X as a binomial random variable with parameters:

$$E[X] = np \Rightarrow \begin{cases} n = \text{round}\left(\frac{E[X]}{0.5}\right) \\ p_X = \frac{E[X]}{n} \end{cases}$$

$$X \sim \text{Bin}(n, p_X)$$

2) Approval or not at an exam: The "pass/not pass an exam" event is modeled as a Bernoulli random variable Y with parameter p_Y. For this implementation, we generate a random number u uniformly distributed between 0 and 1; if $u < p_Y$, the student passes the exam. The value for p_Y depends on the course, since it is extracted from the course's web page as described in Section II-A.

$$Y \sim \text{Bernoulli}(p_Y)$$

3) Grades distribution: The grades of the students for each exam are generated following the distribution extracted from the courses web page. We consider that the number of samples for some courses are smaller, so in the case where one course present at least one mark (from 18 to 30) that was not achieved by any student, the count for all marks are incremented by one. This allows every mark to be achieved slightly changing the distribution.

With the distribution, we are able to compute the cumulative distribution of grades $F(z)$ for each course. For drawing a value from the given distribution, we generate a random value u uniformly distributed between 0 and 1. The grade of the student for the exam will be z, such that $F(z-1) < u <= F(z)$.

4) Rejecting a mark: The main addition to this simulation in comparison to the previous one is the possibility of a student to reject a mark. Again, we model it as a Bernoulli random variable W with parameter p_W. This time, p_W follows a sigmoid, depending on both a rejection threshold r of each student and the mark achieved m, given by

$$p_W(m, r) = \frac{1}{1 + \exp(m - r)}$$

That means that the probability of a student rejecting a mark equals to the rejection threshold is 50%. Moreover, if we consider a rejection threshold $r = 24$, for example, the rejection probability for each mark is as shown in Figure 1. In the case of $m = 30$, the maximum grade possible, we guarantee $p_W = 0$, so that this mark is not rejectable.

C. Input parameters

The considered input for the simulations are:

- List of courses needed to graduate - can be combined with a list with the semester the courses are offered, such that once a year, there is 2 exam appeals for each course on a specific exam session;
- Rejection threshold r - value of the mark to which the rejection probability is 50%. The probability for other marks follows the graph from Figure 1, shifting the points horizontally.
- Average number of exams taken by session ($E[X]$);

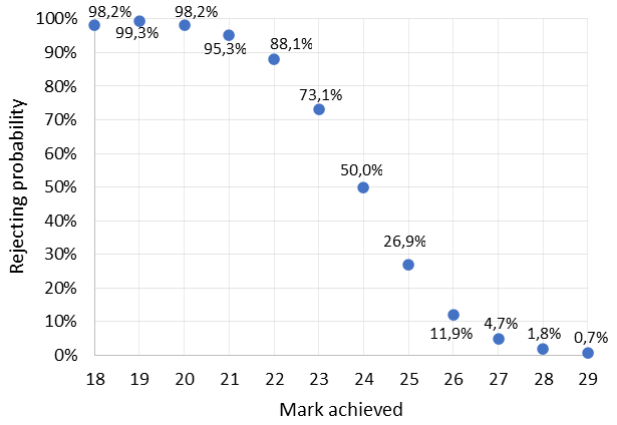

Fig. 1: Probability of rejecting a mark for $r = 24$.

- Exam sessions per year - this allows converting the time to graduate from sessions to years;
- Minimum accuracy for the chosen measures: considering an interval with 98% confidence level;

D. Output measures

There are two output metrics to analyze: the *average grade* of students after graduation and the *average graduation time*, in years.

E. Data structures

During the simulation of a student's career, the main data structures used are:

- `courses`: list containing instances of the `Course` object;
- `graduation_log`: list with tuples (`course`, `semester`, `passed`, `grade`); at each session the student takes the exams on top the list, and at the end of the session it is sorted by `passed`, so that the student only takes exams that he did not get a mark yet.
- `exams_left`: integer initialized as the total number of courses - the simulation stops when it reaches zero;
- `n_sessions`: counter of the total number of sessions needed for the student to graduate.

The final grade will be taken as a simple average from the marks obtained and the graduation time is obtained by dividing the final value of `n_sessions` by the number of sessions per year.

III. Experiments and Results

A. Choice of input parameters

First, we aim to define the reasonable values for the input parameters.

- List of courses: it is expected that the list of courses passed consists of a complete career. We make an exception to the Thesis, that is not implemented.
- Average number of exams taken by session: a good choice for this parameter is a number between 3 and 4, since on average at Politecnico the students take this number of courses each semester. Further, values on a range between 2 and 5 are reasonable, if we consider students that have more difficult to focus on multiple subjects and students who want to graduate faster;
- Rejection threshold r: we may return to Figure 1 to think about it. The probability of rejecting a mark 2 points less than r is almost 90%, while with 2 points more than r, it is around 10%. If we consider a student who does not intend to reject any mark, $r = 0$ is a valid input. On the other hand, a student who intends to graduate with a perfect score may set $r = 50$ and will only accept a mark if it is 30. In between, we may select r in the range 18-25 and cover different approaches of students, towards the possibility of rejecting a mark.
- Exam sessions per year: at Politecnico, there are 3 exam sessions per year, which wil be the value used. The main influence of this parameter is on the chance of the student to take two exam appeals, since they are only available on the first two sessions of the year.
- Accuracy of considered measures: a reasonable value is over 90%, which in our case will be 98%.

B. Data extraction

Following, we analyze the performance of the data collection using the described methods. We make available courses from 3 different possible careers offered by Politecnico, each one with a different method of collection:

- MSc Data Science and Engineering [*data*], collected with a python list of course codes;
- MSc Automotive Engineering (Industrial Processes pathway) [*auto*], collected with a text file with URLs;
- MSc Biomedical Engineering (Bionanotechnologies pathway) [*biom*], collected with a text file with course codes.

C. Validation

For the first experiment, we aim at validating our simulator by analyzing how it performs for different inputs. For example, if we consider a student who does not intend to reject his marks and takes on average 3 exams per session, for each of the considered careers we obtain the following results:

Career	Average final grade	Average graduation time
data	25.22 ± 0.08	2.16 ± 0.04 years
auto	25.37 ± 0.07	2.24 ± 0.04 years
biom	26.58 ± 0.07	2.10 ± 0.04 years

TABLE I: Simulation results for average of 3 exams per session and $r = 0$.

Next we consider $r = 20$, which means a student who may reject some marks, but takes on average 4 exams per session. The results are shown on Table II, where we see that the final grade increased in all courses while the graduation time decreased, since lower grades could now be rejected, but the student took more exams on each session.

Career	Average final grade	Average graduation time
data	26.36 ± 0.06	2.05 ± 0.04 years
auto	26.10 ± 0.05	2.06 ± 0.04 years
biom	27.06 ± 0.06	1.75 ± 0.03 years

TABLE II: Simulation results for average of 4 exams per session and $r = 20$.

From this, we see that not only the trends the output measures follow are as expected, but also their values fall on a reasonable range.

D. Rejection threshold analysis

On this experiment, we fix the input parameters, including the average exams taken per session equals to 3.5, and analyze the behaviour of the average graduation time and the final grade for the considered careers for different values of r. The result for the Msc Data Science and Engineering career is presented in Figure 2.

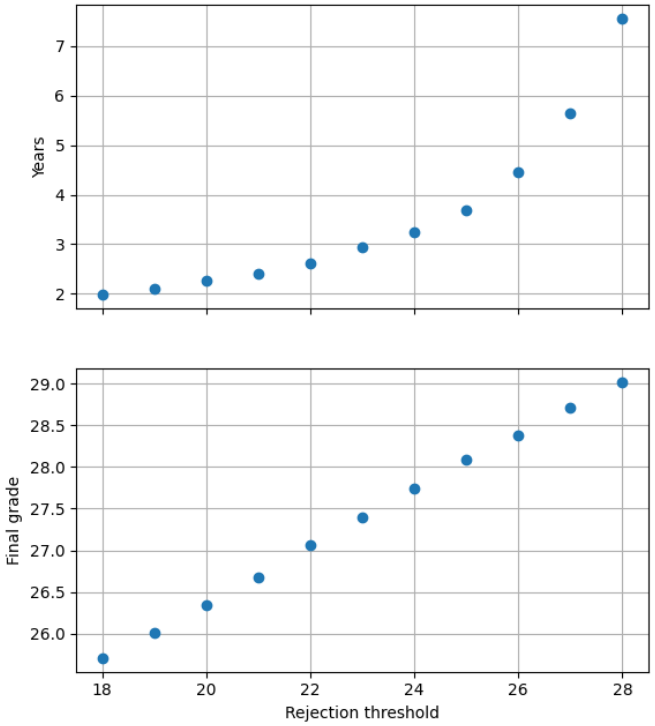

Fig. 2: Graduation time and final grade versus rejection threshold r for MSc Data Science and Engineering.

As shown above, the average graduation time increases exponentially, while the final grade increases linearly as the rejection threshold r increases, i.e. the student rejects lower marks more frequently. For $r = 18$, both final grade and graduation time are not so different in comparison to the previous case of $r = 0$. As for $r = 28$, the student manages to obtain on average a final grade of 29, but for that, it takes more than seven years to graduate. This is an average, that does not take into account the individual aspects of each student, only the generic distribution of grades for the courses. That means that a student who wants to graduate with a final grade of 29, may prepare himself/herself for the exams better than the average students, and by doing that, he/she is able to graduate with the intended final grade in a more suitable time.

E. Average exams taken by session

Finally, we evaluate the graduation time for different values of the average number of exams taken by session $E[X]$. We consider values from 2 to 5 and fix the the rejection threshold $r = 24$. In this case, the variable considered is not directly correlated to the average final grade. It would be the case if the simulator considered the probability of passing an exam and the grades distribution to be different according to the number of exams taken during the session. With the available model, it is more interesting to study only the correlation between the average number of exams taken during each session and the average graduation time of students.

For two of the careers previously loaded, we obtain the results plotted on Figures 3 and 4

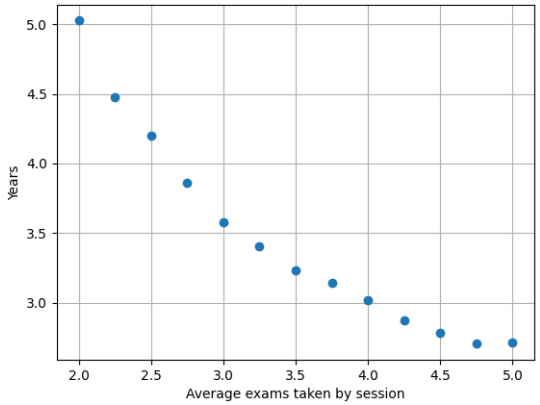

Fig. 3: Graduation time versus average exams taken by session for MSc Data Science and Engineering.

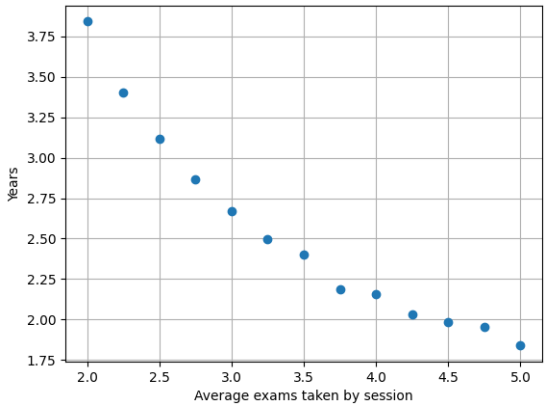

Fig. 4: Graduation time versus average exams taken by session for MSc Biomedical Engineering.

The trends for different careers are similar, as expected. If the students take on average more exams at each exam session with the same chance to be approved as when he takes less exams, the graduation time decreases. In addition, here again we confirm that we were able to individualize the courses and careers by using the data sources, since the actual values differ from one graph to another. That means that, based on the data extracted from the Polito web pages, the approval statistics allow students to graduate sooner or later depending on the career and courses they follow.

IV. Conclusion

The activity was developed in such a way that the simulator previously implemented during Lab G4 could be fine tuned. Now we can extract data from the internet and use it to get individualized information on each course a student takes, besides also allowing they to reject lower marks according to a defined rule.

Finally, some functionalities could still be implemented, for example, students' individual aspects of studies, an approach for the thesis and the different types of examination of each course.