Algorithmic Decision Theory

Lecture 4: Evaluation models
Measure and aggregate performances

R. Bisdorff

FSTC - CSC/ILIAS

10 mars 2020

Content

What is a grade ?

Definition

A grade is an evaluation of the performance of a student in a given course ; an indication to which level a student fulfills the objectives of the course.

Comment

- *A grade should always be interpreted with respect to the objectives of the course.*
- *A grade may have several pedagogical functions such as certifying a certain performance level or being a hint indicating the student's strengths and weaknesses.*
- *A grade is also a public sign addressed to the parents, the University administration, future employers etc.*

The exam types

- Oral or written exams, documents allowed or not,
- Continuous evaluations or single final exam,
- The duration of the exam.

On grading

Grading students copies relies on a number of conventions like :

- Grading scale : 0-20 (France, Belgium & Luxembourg), 0-30 (Italy), 6-1 (Germany), 0-100 (USA), $\{F, E, D, C, B, A\}$ (USA & Asia),
- The model solution giving the repartition of points per question,
- The weight of different exams in the final grade,
- There may be a certain threshold level (10/20 for instance) required in order to validate a course.

Required properties of the grading

- Reliability : For similar copies, the grading should give similar results.
- Faithful validity : the grade given should only measure what was asked for and nothing else.

Empirical properties of the grading

- In mathematics, a difference in grades of 2 points on a $0 - 20$ scale may be commonly observed for similar copies. Motivated grading differences of up to 9 points do occur.
- In 50% of the cases, a second grading by the same corrector leads to a significantly different result than the first one.
- The grades show a high auto-correlation with the apparent level of the student : similar copies from presumably good and presumably weak students commonly obtain dissimilar grades in favour of the good ones.

Empirical properties of the grading – continue

- The order of the copies has an incidence on the grading result. The spread of the grades given by the same corrector commonly augments with time.
- There appear anchorage phenomenas : It is always better to be graded after a weak copy than after an excellent one.
- The overall presentation of a copy –writing, cleanliness – has certainly an influence on the grading result, even if the corrector is supposed to do not care about.

Interpreting grades

- In Europe, grades give generally the impression that they are numerical measures.
- Yet, there is a problem with the minimum grade 0. It does not signify that a student does know nothing !
- There is also a problem with the maximum grade 20. Two excellent students getting 20/20 are not necessarily equivalent !
- What is the genuine scale type of exam grades : ratio, interval, only ordinal ?

Interpreting grades – continue

- If a grading scale is supposed to be of ratio type, all grading differences must in theory be commensurable.
- Yet, very high and low grades for instance do not verify in practice this hypothesis.
- The same is also commonly the case when there exists a validating threshold grade (10/20 for instance). Grading differences, even small, around such a threshold level become consequently more significant : the difference between 10 and 11 is not the same as the one between 18 and 19 for instance.
- Furthermore, grades slightly below the validating threshold are commonly avoided by the correctors.

Interpreting grades – continue

- The preceding problems give arguments to the promoters of Anglo-Saxon alphabetical – i.e. ordinal – grades : generally E or F to A (best grade).
- As a consequence, a large majority of students are often given a neutral grade like B or C.
- In order to better discriminate the effective performances, one introduces qualitative decorations like $+$ and $-$: $B+$ signifying a grade slightly inferior to A, $B-$ a grade slightly better than C.
- It is worthwhile noticing that all these ordinal grades are translating a certain range of number of points or percentages obtained in fact in the underlying exams !
- Finally, one observes that grading differences covering the validating threshold level appear mostly being incommensurable. Consequently, grading scales in general are in fact by essence only more or less ordinal scales.

Rules for aggregating grades

- In order to validate a programme or a degree, it is common usage to aggregate grades obtained in the same and even in different courses.
- Three principles for aggregating are generally used :
  - Conjunctive aggregation
  - Weighted mean
  - Required threshold grades

Conjunctive aggregation

- The students must simply validate all their exams in a given time in order to get their degree.
- Advantage : No commensurability hypothesis concerning the individual grades is required.
- Disadvantages :
  - Many students risk to eventually fail their degree.
  - There are only two types of results : valid and invalid.
  - No formative results may be expressed : slightly insufficient for example in order to not discourage and positively stimulate a student to enhance his performance for instance.
  - No distinction can be expressed : The students are not stimulated towards giving their best.

Weighted mean

- Often, aggregating grades is done by a simple weighted average of individual grades obtained in each course.
- To validate a study programme or degree, this weighted average grade is then compared to standard values like 10/20, or 14/20, 16/20 etc. to attribute a distinction.
- The weighted average requires, contrary to the conjunctive aggregation, the full compensation between all possible grades.

Validating threshold levels

- Required minimal thresholds for validating a course or a whole programme are commonly introduced in order to avoid full compensation between individual grades (a 0/20 grade being compensated by a 20/20 grade for instance).
- Sometimes, the average grade has to reach a certain level (14/20) before compensating is allowed.
- Commonly, all three principles may be combined in practice.

Weighted average grade : Notations

Definition

- We suppose that all grades are expressed on a $0 - 20$ scale.
- We denote $g_i(a)$ the grade obtained by a student a in the course i ($i = 1$ to n).
- We denote w_i the (strictly positive) weight allocated to course i in the evaluation of the final grade.
- The final grade $g(a)$ of student a is computed as follows :

$$\boxed{g(a) = \sum_{i=1}^{n} w_i \cdot g_i(a)}$$

Weighted average grade – continue

Comment

- *The weights w_i are commonly expressed as integer numbers (number of lectures, hours, lessons, or ECTS ...).*
- *The weights w_i may always be normalised without loss of generality as follows :*

$$w_i' = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

- *Normalised weights w_i' – rational numbers – are thus confined between 0 and 1 and $\sum_{i=1}^{n} w_i' = 1$.*
- *The average grade, computed with normalised weights, will be expressed on the same scale ($0 - 20$ for instance) as the individual courses' grades.*

Methodological problems

Example (1. An undesirable effect of the compensation)

Consider four students $\{a, b, c, d\}$ enrolled in a study programme consisting of two courses $\{g_1, g_2\}$ of same weight and where they have obtained the following grades :

	g_1	g_2
a	11	11
b	5	19
c	20	4
d	4	6

Student a shows satisfactory results in both courses, whereas student d shows very weak results. On the contrary, b and c are both excellent students in one course and weak in the other. Globally, a should be ranked before b and c, and both ranked again before d

Example (1) – continue

Comment

Aggregating the four students grades with a weighted average results in following figures :

	g
b	*12*
c	*12*
a	*11*
d	*5*

Students b and c are ranked before student a. One may even verify that no other weighting of the two courses will allow to rank a before b and c ! Use a weighted average is in fact incompatible with the idea of promoting those students that do reasonably good in all courses.

Exercise(s) (1. An undesirable effect of the compensation)

Show that, when aggregating with a weighted average the grades above, there does not exist any possible weighting of both courses such that a is ranked before b and c

Methodological problems – continue

Comment

Practical consequences of unlimited compensation :

- *Using a weighted average as rule for aggregating grades may turn students towards concentrating their efforts on a limited number of courses only by relying on the compensation mechanism for getting a sufficient final grade.*
- *Requiring minimal threshold grades may limit, but not completely inhibit, this undesirable effect.*

Methodological problems – continue

Example (2. Interactions between performances to aggregate ?)

Consider four students $\{a, b, c, d\}$ enrolled in a programme consisting in statistics (S), mathematics (M) and economics (E). They got the following grades :

	g_S	g_M	g_E
a	18	12	6
b	18	7	11
c	5	17	8
d	5	12	13

Student a should be ranked before student b in an engineering study programme. b is, even more, weak in maths, which is convenient neither for an engineering nor an economics degree. With a similar reasoning, d is much better than c when considering an economics degree.

Methodological problems – continue

Comment

Interactions between performances :

- *Whereas the preceding rankings seam quite reasonable, they are however not compatible with the weighted average rule.*
- *When the statistics results are excellent, the weight of mathematics outranks the one of economics (a outranks b).*
- *However, showing weak grades in statistics leads to consider that the weight of economics outranks the one of mathematics (d outranks c)*
- *These interactions between course subjects, despite the fact of being quite common in practice, are not compatible with the weighted average rule.*

Methodological problems – continue

Example (3. Incommensurable differences between grades ?)

Consider two students enrolled in a programme with two courses of same weight. The grading is done on a $0-20$ scale and a final grade of at least 10 is required in order to validate the programme.

	g_1	g_2
a	11	10
b	12	9

Both students obtain the same average grade 10.5 and validate equivalently the programme. The difference between 12 and 11 in the first course exactly compensates the difference between 10 and 9 shown in the second course.

Methodological problems – continue

Comment

Incommensurable differences between grades :

- *As 10 is the threshold for validating the programme, one may suppose that the difference observed in the first course is more important than that observed in the second one.*
- *Consequently, student a must in fact have better validated the programme than student b ?*
- *Indeed, a was conjointly successful in both courses, whereas b failed one of the two courses.*
- *With the weighted average rule, a difference of one point is required to have uniformly the same signification all along the scale.*

Methodological problems – continue

Example (4. Incommensurable differences between grades ?)

Reconsider the three students enrolled in the same programme as in Example (3) :

	g_1	g_2
a	$14-x$	$14+x$
b	14	14
c	$14+x$	$14-x$

Comment

The three students obtain the same average of 14 (for $x = 1, 2, ..., 5$) and validate equivalently the programme with a final grade 14 (good).

If $x = 1$, this result is acceptable.

If $x = 5$, this result is no more acceptable.

How to aggregate ordinal grades ?

Example (5. grading on an ordinal scale)

Consider three students enrolled in a study programme consisting of three courses graded from 0 to 20 points and where a grade of 10/20 is required for succeeding the programme. If the grading scale is purely ordinal, the following grades will show the same result for each student.

	g_1	g_2	g_3
a	12	5	13
b	13	12	5
c	5	13	12

	g_1	g_2	g_3
a	11	4	12
b	13	13	6
c	4	14	11

In the first case, all three students validate, whereas, in the second case, only b validates the programme.

How to aggregate ordinal grades

Example (6. The US Grade Point Average GPA)

As the courses are graded on alphabetical levels from E to A, one has to numerically encode these levels. A common conversion schema is the following :

level	grade	mention
A	4	(excellent)
B	3	(very good)
C	2	(good)
D	1	(satisfactory)
E	0	(failure)

Comment

- *The choice of grades 4 to 0 is arbitrary.*
- *A constant difference between two adjacent levels is assumed.*
- *Obtaining an excellent level A is supposed to be 4 times as performing as obtaining as satisfactory level D ! ? !*

Example (6) Computing the GPA – continue

Exams in the US are generally graded from 0 to 100 %. Suppose that three student obtained the following grades in three courses :

	g_1	g_2	g_3
a	90	69	70
b	79	79	89
c	100	70	69

Conversion schema :

level	interval	grade
A	90 – 100%	4
B	80 – 89%	3
C	70 – 79%	2
D	60 – 69%	1
E	0 – 59%	0

Example (6) Computing the GPA – continue

Converting the results :

	g_1	g_2	g_3
a	A	D	C
b	C	C	B
c	A	C	D

Computing the GPA :

	g_1	g_2	g_3	GPA
a	4	1	2	2.33
b	2	2	3	2.33
c	4	2	1	2.33

Comment

All three students obtain the same GPA value 2.33.

Example (6) Computing the GPA – continue

Other conversion schema :

level	interval	grade
$A+$	98 – 100%	10
A	94 – 97%	9
$A-$	90 – 93%	8
$B+$	87 – 89%	7
B	83 – 86%	6
$B-$	80 – 82%	5
$C+$	77 – 79%	4
C	73 – 76%	3
$C-$	70 – 72%	2
D	60 – 69%	1
E	0 – 59%	0

Conversion results :

	g_1	g_2	g_3
a	$A-$	D	$C-$
b	$C+$	$C+$	$B+$
c	$A+$	$C-$	D

Computing the GPA :

	g_1	g_2	g_3	GPA
a	8	1	2	3.66
b	4	4	7	5.00
c	10	2	1	4.33

Student b obtains now clearly a better result.

Aggregating ordinal performances

Example (Condorcet's method)

Consider three students enrolled in a study programme consisting in three courses of same weight and who obtained the grades shown here :

	g_1	g_2	g_3
a	13	12	11
b	11	13	12
c	14	10	12

Comment

- *The three students obtain the same average grade* 12.
- *Consider now that a difference of one point on the grading scale is not really significant for warranting an effective performance difference.*
- *Student a shows at least as good grades as b and c in all the courses.*
- *However, students b are c are only in two out of three courses at least as good as student a.*

Exercise(s)

Here the table of grades obtained by four students : a, b, c, and d, in five courses : C_1, C_2, C_3, C_4 *and* C_5.

course	C_1	C_2	C_3	C_4	C_5
ECTS	2	3	4	2	4
a	11	13	9	15	11
b	12	9	13	10	13
c	8	11	14	12	14
d	15	10	12	8	13

An award is granted to the best amongst these four students.

1. *Who would you nominate ?*
2. *Explain and motivate your selection algorithm.*

Exercise(s) (Random students performance tableaux)

1. *Use the `Digraph3` Python resources for generating realistic random students performance tableaux (see the `randomPerfTabs.py` module).*
2. *Design and implement a* **fair diploma validation decision rule** *based on the results obtained in 9 weighted Courses.*
3. *Run simulation tests with random students performance tableaux for validating your design and implementation.*

Concluding

- Grading accurately someones performances is generally a difficult task in practice.
- Grading procedures are in general quite complex and must not be seen as simple as physical weight, time and length measures.
- Aggregating grades needs taking into account potential imprecision, uncertainty as well as known cognitive biases.
- Aggregating rules have to be analyzed with great attention. The simplests and evidents do not necessarily give the expected results.