Milestone: the Weights of Different Factors on Influencing Average Grades

Kewen Ding R0733305

Rui Zhu R0734658

Motivation:

We want to see whether there is a connection between possible features and the average grades of a student. Except for alcohol consumption, we would also like to see whether other factors play a role here. From the last feedback, our coach suggested we could use the feature selection.

Method

Gradient Descent and Normal Equation will be mainly focused on for now. We will go through all the features together, analyzing if this single feature can fit the data well. To do so, Gradient Descent and Normal Equation are used respectively to determine the weight (theta) for each feature. After that, all the weights will be sorted from the largest to the smallest. As a result, a general idea of which features are of great importance can be obtained.

Firstly, two attributes are chosen -- Dalc and Walc (Workday alcohol consumption and weekend alcohol consumption). These two attributes range from 1 -very low to 5- very high.

The output has three attributes G1, G2, and G3, which means first-period grade, second-period grade, and final grade, They range from 0 to 20. We will take the mean value for now.

Preliminary Result

We first analyzed the data by linear regression Gradient Descent. The result is shown in the first two columns. Because age is about 15 -18, while other attributes are among 1 to 5, the cost function decreases slowly without feature normalization. For the first two columns, the cost function decreases very slowly. 60000 times of Gradient Descent did not give satisfying theta. Till the 100000 times, it gives the appropriate theta. From the cost of each method, it has a relatively higher cost than that of others. With the normal equation, the smallest cost and a better theta can be obtained.

From the result, it can be seen that the alcohol consumption does have an impact on students' scores, and the Dalc (-0.26)(Workday alcohol consumption) has a more negative impact on grades than that of Walc(-0.09)(Weekday alcohol consumption). It's reasonable. For all the attributes, the higherPursuit(wants to take higher education), study time (weekly study time), and internet(Internet access at home) rank the highest theta of determining the scores. If the student wants to acquire a higher degree, he or she will study longer and get a better grade. Other attributes like access to the internet, Mother's education, and whether has extra-curricular activities are also crucial.

The factors with the most negative impact on student's scores are failures(number of past class failures), school support(extra educational support), and paid(extra paid classes within the course subject). Because his or her grades are not satisfying, the parents and teacher may suggest him to go to an extra paid class. Therefore, the paid class shows a negative weight in the prediction.

Next step

For the future work, firstly, all the 30 features will be taken into consideration since for now we still have some problems applying different types of data into our project, but it will be solved before long.

Neural networks will be applied into our project. Because now we just got a general idea of which features play important roles in our prediction, but they are still separated from each other. Neural networks can really combine these features and come up with a better algorithm that can predict more accurately. We also apply other more advanced techniques if they are taught in the future.

On top of that, considering we have two different tables for the grades of math and Portuguese language, what we can do is use the model obtained from the dataset of the Portuguese language to predict that of the math course and compare it with the real grades to see if our model has a more generalized application.

Appendix

Table1. The theta by different methods(α=0.001)

θ	Unnormalized features (60000 iterations)	Unnormalized features (100000 iterations)	Normalized age(10000 iterations)	Normal Equation
theta0Const	1.8071	2.4034	7.7051	8.3121
Higher Pursuit	2.1133	2.0846	2.1066	1.7307
studyTime	0.5166	0.5142	0.5966	0.4946
internet	0.4759	0.4747	0.5486	0.4639
activities	0.3005	0.2976	0.2414	0.2677
MotherEdu	0.2565	0.2532	0.2596	0.2218
FatherEdu	0.2097	0.2076	0.2093	0.1862
age	0.4012	0.3731	0.1206	0.0973
Family relationship	0.1684	0.1617	0.2194	0.095
Family support	0.0663	0.0589	0.0008	-0.0173
absences	-0.0266	-0.0266	-0.0175	-0.0264
goOutActivity	-0.0561	-0.0543	-0.0134	-0.037
Weekend alcohol consumption	-0.0595	-0.0627	-0.0581	-0.0931
health	-0.0903	-0.0926	-0.074	-0.1158
Free time	-0.0553	-0.0623	-0.0618	-0.1294
traveltime	-0.0878	-0.0971	-0.0523	-0.1852
nursery	-0.1142	-0.121	-0.0348	-0.1908

Weekday alcohol consumption	-0.2857	-0.2834	-0.2422	-0.2664
romantic	-0.4321	-0.4279	-0.3112	-0.387
paid	-0.6522	-0.6616	-0.6924	-0.6447
School support	-0.7302	-0.7519	-0.8901	-0.9348
failures	-1.3902	-1.3777	-1.1802	-1.2726

Table2. The cost function of different methods(α=0.001)

	Unnormolized feature 60000	Unnormolized feature 100000	Normalized age	Normal Equation
cost	2.767622	2.751127	2.731991	2.6948

Table3. Annotation of different attributes

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - 1 hour)

14. studytime - weekly study time (numeric: 1 - 10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:

1. G1 - first period grade (numeric: from 0 to 20)
2. G2 - second period grade (numeric: from 0 to 20)
3. G3 - final grade (numeric: from 0 to 20, output target)