MS&E 226 Project part 1

Kei Masuda, Erin Brown

TOTAL POINTS

9 / 10

QUESTION 1

1 Project Part 1 9 / 10

✓ **+ 6 pts Minimum requirements: completely met**

+ **4 pts** Minimum requirements: missing one or two aspects

+ **2 pts** Minimum requirements: missing several aspects

+ **0 pts** Minimum requirements: not met overall

+ **4 pts** Subjective evaluation: extremely thoughtful analysis and very well-written/organized report

✓ **+ 3 pts Subjective evaluation: thoughtful analysis and well-written/organized report**

+ **2 pts** Subjective evaluation: acceptable report. Went beyond minimum requirements in one or two aspects, e.g., data cleaning or data visualizations/transformations, but overall depth of analysis could be improved.

+ **1 pts** Subjective evaluation: did not go beyond minimum requirements. Many improvements are possible.

① How are you calculating distance between covariate vectors? It is important to consider the range/scale of the covariates when defining a notion of distance.

② For additional transformations that you might consider, see here: https://rcompanion.org/handbook/I_12.html

③ Just FYI, to see if there is correlation between two categorical variables, you can use the Pearson chi-square test.

To see if there is correlation between a categorical variable and a continuous variable, you use a one-way ANOVA test.

The first answer here (https://datascience.stackexchange.com/questions/893/how-to-get-correlation-between-two-categorical-variable-and-a-categorical-variab) gives a great description of these methods with examples and R code - please take a look. Basically, you run hypothesis tests to determine whether or not the two variables are independent (if they are not independent, then there is some correlation between them). The null hypothesis in these hypothesis tests is that of independence (meaning, the two variables are independent, and thus, uncorrelated). (Actually, to be completely accurate, the null hypothesis in the one-way ANOVA test is essentially that of independence (specifically, the null hypothesis is that there is no difference in the means of the continuous variable, among the different categories of the categorical variable).) Anyhow, smaller p-values provide evidence supporting rejection of the null hypothesis, and thus, small p-values are indicative of low correlation.

④ How did you set the regularization parameter for ridge and lasso regression? As a reminder, generally, you want to use a data-driven method like cross-validation to set the regularization parameter!

⑤ These are nice questions!

⑥ You might also consider regularized logistic regression using lasso. Just like for lasso linear regression, you can also use the glmnet() function in R.

⑦ It's quite important to set the regularization term C accurately. If you're encountering numerical optimization issues, you might use one of the R packages, rather than design your own solver.

⑧

Yes, this is not that big of a deal.

1 Project Part 1 9 / 10

✓ + **6 pts** **Minimum requirements: completely met**

+ **4 pts** Minimum requirements: missing one or two aspects

+ **2 pts** Minimum requirements: missing several aspects

+ **0 pts** Minimum requirements: not met overall

+ **4 pts** Subjective evaluation: extremely thoughtful analysis and very well-written/organized report

✓ + **3 pts** **Subjective evaluation: thoughtful analysis and well-written/organized report**

+ **2 pts** Subjective evaluation: acceptable report. Went beyond minimum requirements in one or two aspects, e.g., data cleaning or data visualizations/transformations, but overall depth of analysis could be improved.

+ **1 pts** Subjective evaluation: did not go beyond minimum requirements. Many improvements are possible.

1 How are you calculating distance between covariate vectors? It is important to consider the range/scale of the covariates when defining a notion of distance.

2 For additional transformations that you might consider, see here: https://rcompanion.org/handbook/I_12.html

3 Just FYI, to see if there is correlation between two categorical variables, you can use the Pearson chi-square test.

To see if there is correlation between a categorical variable and a continuous variable, you use a one-way ANOVA test.

The first answer here (https://datascience.stackexchange.com/questions/893/how-to-get-correlation-between-two-categorical-variable-and-a-categorical-variab) gives a great description of these methods with examples and R code - please take a look. Basically, you run hypothesis tests to determine whether or not the two variables are independent (if they are not independent, then there is some correlation between them). The null hypothesis in these hypothesis tests is that of independence (meaning, the two variables are independent, and thus, uncorrelated). (Actually, to be completely accurate, the null hypothesis in the one-way ANOVA test is essentially that of independence (specifically, the null hypothesis is that there is no difference in the means of the continuous variable, among the different categories of the categorical variable).) Anyhow, smaller p-values provide evidence supporting rejection of the null hypothesis, and thus, small p-values are indicative of low correlation.

4 How did you set the regularization parameter for ridge and lasso regression? As a reminder, generally, you want to use a data-driven method like cross-validation to set the regularization parameter!

5 These are nice questions!

6 You might also consider regularized logistic regression using lasso. Just like for lasso linear regression, you can also use the glmnet() function in R.

7 It's quite important to set the regularization term C accurately. If you're encountering numerical optimization issues, you might use one of the R packages, rather than design your own solver.

8 Yes, this is not that big of a deal.