IMT573 Lab 8: validate model

December 2, 2020

Instructions

In this lab you use logistic regression, add random data to Titanic dataset, and explore how the model performance changes on the training and validation data.

1 Add random number to Titanic data

We use the same Titanic data you have already seen several times. But now we add random numbers to this dataset and see how the model performance changes on both training and validation data.

1. Load Titanic data. We preserve *survived*, *age*, *pclass*, and *sex*, *sibsp*, *parch*, and *fare*, and remove the other variables. (The other variables either contain too many missings, are are distinct, like names.)
2. Remove all the rows that contain missings. You can do it like this:

```
df <- df[complete.cases(df),]
```

3. Now do the following:
   (a) Split data into training/validation parts (80/20)
   (b) Train logistic regression model on training data using all the variables.
   Hint: you can tell the model "all variables" using the formula "`survived ~.`"
   (c) Compute and report accuracy on both training and validation data
4. Next, add 500 columns of random numbers to the titanic data, in a similar fashion as what we did in the class. Check using `dim` that it's dimension now is 507 columns.

   Hint: you can create a 100×50 random matrix by:

```
R = matrix(rnorm(100*50), nrow = 100)
dim(R)

## [1] 100  50
```

5. Now select first $\mathsf{N} = 2$ columns from your extended titanic dataframe. Repeat the tasks in 3.
6. Finally, let your N grow slowly and observe what happens to the training and validation accuracy.

   Warning: the computations get very slow if N is large. Choose initially $\mathsf{N} < 50$, and try larger values only when you know the code works.

2 Challenge (not graded)

If you have time and interest, then consider also making a plot where you have number of columns N on the horizontal axis, and training/validation accuracy on the vertical axis (plot two lines with different color).

Comment how do the lines perform as N grows.