Foundations of Computational Math – Final Exam Part 3

Sam Reeves

11/24/2021

Part 3. Kaggle Competition – House Prices: Advanced Regression Techniques

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
house_train <- tibble(read.csv('prob3/train.csv'))
house_test <- tibble(read.csv('prob3/test.csv'))
```

Descriptive and Inferential Statistics (5)

> Provide univariate descriptive statistics and appropriate plots for the training data set. Provide a scatterplot for at least two of the independent variables and the dependent variable. Derive a correlation matrix for any three quantitative variables in the dataset. Test the hypothesis that the correlations between each pairwise set of variables is 0 and provide an 80% confidence interval. Discuss the meaning of your analysis. Would you be worried about familywise error? Why or why not?

Linear Algebra and Correlation (5)

> Invert your correlation matrix from above. (This is known as the precision matrix and contains variance inflation factors on the diagonal.) Multiply the correlation matrix by the precision matrix, and then multiply the precision matrix by the correlation matrix. Conduct LU decomposition on the matrix.

Calculus-Based Probability and Statistics (10)

Many times it makes sense to fit a closed form distribution to data. Select a variable in the Kaggle.com training dataset that is skewed to the right, shift it so that the minimum value is absolutely above zero if necessary. Then load the *MASS* package and run *fitdistr* to fit an exponential probability density function. Find the optimal value of λ for this distribution, and then take 1000 samples from this exponential distribution using this value (e.g., rexp(1000, λ)). Plot a histogram and compare it with a histogram of your original variable. Using the exponential PDF, find the 5^th and 95^th percentiles using the CDF. Also, generate a 95% confidence interval from the empirical data, assuming normality. Finally, provide the empirical 5^th percentile of the data. Discuss.

Modeling (10)

Build some type of *multiple* regression model and submit your model to the competition board. Provide your complete model summary and results with analysis. Report your Kaggle.com username and score, and provide a screenshot.