HW3 — Diamond Mining

CS249 — Spring 2016 — D.S. Parker

Assignment Overview

The goal of this assignment is for you to develop models for the diamonds dataset included in the ggplot2 package.
This is a very simple assignment: you are asked to build four models: LDA or QDA, simple Linear Regression, log-scaled Linear Regression, and Logistic Regression. You then just upload the formulas (R commands) you used to construct these models to CCLE.

Step 0: build the numeric.diamonds dataset

This notebook includes commands for buiding a dataset called `numeric.diamonds` that you are to use for this assignment.
The diamonds dataset has 3 categorical attributes (cut, color, clarity) that are ordered. The numeric.diamonds dataset is a numeric version of the diamonds dataset in which all these categorical attributes are converted to integer codes.
For example, there are 7 colors, with the ordering: J < I < H < G < F < E < D (J is worst, D is best). We implement these by replacing J with the value 1, I with the value 2, . . . , and D with the value 7.
After doing this transformation for cut and clarity also, the result is an entirely numeric dataset called numeric.diamonds.
In addition to this notebook, we've provided another called Diamonds.ipynb for gaining intuition about the data by walking through some exploratory graphics. Many aspects of the dataset are displayed. You do not have to use this notebook, it is totally optional, but it is included since visualization can help.

Step 1: build a training set and test set (as subsets of numeric.diamonds) – using your UID

First, set the random number generator seed to your UID. Then generate a training set and test set using the following commands:

```
MY_UID = 123456789    ########## you must enter your UCLA UID here !!!
set.seed( MY_UID )

n = nrow( numeric.diamonds )
sample.size = 0.75 * n
training.row.ids = sample( (1:n), sample.size )

my.training.set = numeric.diamonds[  training.row.ids, ]
my.test.set  = numeric.diamonds[ -training.row.ids, ]   # set complement of training.set.ids
```

Please use exactly these commands to construct your training set and test set. Also, use the training set to construct each model, and use the test set to compute the accuracy of each model. The grading program will re-compute your model and its accuracy using this method.

Step 2: Evaluate the accuracy of four Baseline Models about diamonds

These are the baseline models:

```
qda( cut ~ .,           data=my.training.set )
lm(  price ~ .,         data=my.training.set )
lm(  log10(price) ~ .,  data=my.training.set )
glm( I(price>1500) ~ ., data=my.training.set, family=binomial )
```

You are to develop a notebook that computes the accuracy of each of these models on `my.test.set`.

Step 4: generate improvements on the Baseline Models

For the numeric.diamonds dataset you are to develop a notebook that evaluates accuracy of four baseline models in R:

1. a LDA or QDA classification model that predicts a diamond's Cut.
2. a linear regression model that predicts a diamond's Price.
3. a linear regression model that predicts a diamond's log10(Price).
4. a logistic regression model that predicts whether a diamond's Price is above $1500.

As an example, you might produce these models:

1. `qda( cut ~ price + table + color + clarity, data=my.training.set )`
2. `lm( price ~ carat + x + y + z + clarity, data=my.training.set )`
3. `lm( log10(price) ~ table + log10(carat) + color, data=my.training.set )`
4. `glm( abs(price>1500) ~ carat + table + clarity, data=my.training.set, family = binomial )`

As these examples show, details matter: you must specify the complete formula for each model in detail, listing all variables included.
Please choose attributes so that your model outperforms the baseline model on `my.test.set`.

Step 4: generate a CSV file `HW3_output.csv` including your 4 models

If these were your four models, then to complete the assignment you would create a CSV file `HW3_output.csv` containing eight lines:

```
33.333, qda( cut ~ .,           data=my.training.set )
88.888, lm(  price ~ .,         data=my.training.set )
77.777, lm(  log10(price) ~ .,  data=my.training.set )
88.888, glm( I(price>1500) ~ ., data=my.training.set, family=binomial )
44.444, qda( cut ~ price + table + color + clarity, data=my.training.set )
99.999, lm( price ~ carat + x + y + z + clarity, data=my.training.set )
99.999, lm( log10(price) ~ table + log10(carat) + color, data=my.training.set )
99.999, glm( I(price>1500) ~ carat + table + clarity, data=my.training.set, family=binomial )
```

The first four lines show the four Baseline Models. The next four are your models that improve on the Baseline model.
Each line gives **the accuracy of a model on** `my.test.set`, as well as **the exact command you used to generate the model**. You must develop procedures for computing accuracy that follow instructions in the notebook.
For each of the baseline models, you will get 18 points for computing accuracy values correctly. You will get 7 points for producing a model with higher accuracy. Thus the maximum possible for this assignment is 100 points.

Step 5: upload your CSV file and notebook to CCLE

Finally, go to CCLE and upload:

- your output CSV file `HW3_output.csv`
- your notebook file `HW3_Diamond_Mining.ipynb`

We are not planning to run any of the uploaded notebooks. However, your notebook should have the commands you used in developing your models — in order to show your work. As announced, all assignment grading in this course will be automated, and the notebook is needed in order to check results of the grading program.