Assignment 1: Linear Regression

Deadline: Monday, April 16, 2012

This assignment will contribute 7 points to your final grade.

Description

In this assignment, you will fit a linear regression model to a wine quality dataset (available from the UCI Machine Learning repository\(^1\)). The dataset consists of two files—one for red wine (*winequality-red.csv*) and one for white wine (*winequality-white.csv*). Read the description in the UCI link to learn more about the source and organization of the datasets. The files are provided in the assignment page for your convenience.

In brief, the data contains different wine characteristics (e.g. pH, alcohol, density, etc.) and the corresponding quality level assigned by a human. A linear regression model can be learned on the data to predict the quality, given values for the wine features.

Your objective is to determine if learning separate models on both wine types is necessary, or a single model is sufficient. In addition, you will determine what characteristics (three most important) about a wine plays the most prominent role in determining its quality. The remainder of the document describes how this study is to be pursued.

Task 1: Writing an experiment

Begin your study by importing the data into R. The next steps will involve performing multiple experiments on the data. Let's define what we mean by an experiment. The inputs to an experiment are a training set and a test set (or a data set with a number `frac` indicating what fraction to use as training data). The steps involve:

1. Divide data set into training and test data using the indicated fraction `frac` (only if not explicitly provided)

2. Extract and keep the “quality” value as a separate column vector (for both training and test data); you may also pass the dependent variable data separately

3. Standardize the training and test data (remember that test data is standardized with statistics from training data)

4. Add bias column to training and test data

5. Fit a linear model to the training data

6. Compute root-mean-squared-error (RMSE) on both training and test data

   (a) wine quality is indicated by a value from 0 to 10 (integer)

   (b) values predicted by the model outside these limits should be corrected: negative numbers become zero, and all predictions above 10 become 10.

   (c) fractional values can be converted to an integer using `floor`, `round`, or the `ceiling` functions (you will get an RMSE value corresponding to each function)

---

\(^1\)http://archive.ics.uci.edu/ml/datasets/Wine+Quality
7. Return as a R list:

(a) weights learned
(b) RMSE on training data (three of these corresponding to the three conversion functions)
(c) RMSE on test data (three of these corresponding to the three conversion functions)
(d) predicted values on training data
(e) predicted values on test data

Writing this process as a function will greatly speed up the rest of the tasks. Depending on how you structure your R code, feel free to return any other value, in addition to those asked for in item 7.

**Task 2: Selecting a conversion function**

Using the red wine data, perform 100 experiments each with \( \text{frac} = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 \) and 0.9. For each of the three functions, note the average RMSE values you see across all these experiments (900 in total). Repeat with the white wine data. Pick one function based on the observed errors. Discuss in your report what these RMSE values mean and the reason behind your selection. Use your selected function in the following tasks.

**Task 3: Deciding a model**

You have multiple options while picking a model to predict wine quality. Here are some choices:

1. Fit on red wine data, use for prediction on red wine; similarly, fit on white wine data, use for prediction on white wine
2. Fit on red wine data, use for prediction on any wine; similarly, fit on white wine data, use for prediction on any wine
3. Fit on a mix of red and white wine data, use for prediction on any

Your final selection will be based on the errors that you observe. You have already computed the errors for Case 1 in Task 2. For the other two cases:

**Case 2:** Run an experiment using the entire red wine data as training dataset, and the entire white wine data as testing dataset. Similarly, run another experiment using the entire white wine data as training dataset, and the entire red wine data as testing dataset.

**Case 3:** Run 100 experiments each with \( \text{frac} = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 \) and 0.9, using the red+white wine data as a data set. Note that there are more instances of white wine than that of red wine. The training data you create should have \( \text{frac} \) representation of both types, i.e. a combination of \( \text{frac} \) fraction of the red wine data and \( \text{frac} \) fraction of the white wine data will be the training data. As before, error values are averaged across the 900 experiments. In this case, while computing the error on the test data, compute two test error values—(i) on those test instances that came from the red wine data, and (ii) on those test instances that came from the white wine data.

After completing these experiments, you should have six different error values, one each from the R-on-R, W-on-W, R-on-W, W-on-R, M-on-R and M-on-W studies. Here \( R=\text{Red}, W=\text{White}, M=\text{Mix} \) and \( \text{Tr-on-Ts} \) means you learned the model on \( \text{Tr} \) data and used it to predict on \( \text{Ts} \) data.

Create an informative barplot using these error values. Discuss whether you would prefer to have one model for each wine type, or use the model learned from the mix data. Your conclusions should be based on the error values (not personal preferences). You should write how you reached the conclusion.
Task 4: Top three features

Use the model that learns on mix data for this task. Run 100 experiments, each time with a randomly chosen value for \(\text{frac}\), using the red+white wine data as a data set. Note the top three attributes (in terms of highest absolute weight), excluding the \(\text{bias}\). Keep track of how many times an attribute makes it to the top three. After the experiments, create an informative barplot showing the percentage of times a certain attribute was chosen as one of the top three. Do the top three attributes depend heavily on the train/test split? Discuss your observations in this matter.

Report

Your report is **NOT** a collection of statements corresponding to the discussion asked from you. It must be complete in itself, primarily including the problem you are solving, description of the data, the method employed, relevant R code, the experiments performed, the results you observed, informative (with labels, captions, and the remaining glitter!) plots, and most importantly, a discussion along the lines of what is asked in the tasks. You are strongly encouraged to include your own observations and discussions outside of what is asked. Tell me a story!

Grading

This assignment will be graded out of 70 points. The points split is as follows:

<table>
<thead>
<tr>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

The points you receive in each task will depend on how accurately you complete the task. In addition, your R code should be clean and well-commented. Sticking to directions is important! Additional explorations (outside of what is asked in the tasks) is most welcome. The report will be graded based on content, structure, formatting (I do not want to see half of a page empty because the plot does not fit there — reorganize!!), spelling, grammar, and other usual characteristics of a good scholarly document.

Submission

Submit the pdf report (not the Latex or Lyx source) and the R code into Blackboard by **Monday, 16 April 2012, 11:59 PM**. Refer to the late policy on the course website.

You must work alone on this assignment.