**Homework Part 3**

**Linear and Non-Linear Logistic Regression with Multiple Features**

INSTRUCTIONS

Please capture all your work in an Orange file (with extension .ows), a Word file and a handful of Excel files. Please copy the images you generate in Orange into the Word file to illustrate your answers as needed.

IMPORTANT NOTE: If you use certain graphs and/or tables to answer a question or draw a conclusion please include the respective graphs / tables. Do so even if including that graph / table is not specifically asked for in the question.

ASSIGNMENT

Use Orange to perform a multi-feature logistic regression on the forged-bank-notes.xlsx dataset.

In this homework assignment, you will separate genuine from counterfeit bank notes using logistic regression. You’ll train various logistic regression models using your training data and then use your test dataset to generate predictions for each logistic regression model. You’ll compare these predictions more formally by using a widget called Confusion Matrix.

Specifically, carry out the following steps:

1. The dataset has 4 features and a target variable called ‘Genuine’. Use Orange’s File widget to make sure the features are named A1, A2, A3, A4 (these are all numeric features) and the target is named ‘Genuine’. (Note: This is just a step to take and there are no questions to answer.)
2. Use the Feature Constructor widget to create the following features to the dataset. (Note: This is just a step to take and there are no questions to answer.)
   1. A1\*A2
   2. A1\*A3
   3. A3^2 (the square of A3)
   4. A4^3 (the cube of A4)
3. Your dataset should now contain the following 8 features: A1, A2, A3, A4, A1\*A2, A1\*A3, A3^2, and A4^3. Look at the range of values of each feature. The range of values of each feature are sufficiently different. So use the Preprocess widget to normalize the dataset (use the defaults of “Center by Mean” and “Scale by SD”). (Note: This is just a step to take and there are no questions to answer.)
4. Now split the dataset (the one you’ve constructed with all 8 features) into a Training dataset containing 80% of the data and a Test dataset containing the remaining 20% of the data. (Note: This is just a step to take and there are no questions to answer.)
5. Build logistic regression models for 3 different variations of the data set:
   1. Data Set 1: Use the features A1, A2, A3, and A4.
   2. Data Set 2: Use the features A1\*A2, A1\*A3, A3^2, and A4^3.
   3. Data Set 3: Use all 8 features.
6. Use the Predictions and the Confusion Matrix widgets to get the predictions of each model *for the test data set.* Save the predictions from the Predictions widget to an Excel file. Using the information from that file and the information from the confusion matrices, answer the following questions:
   1. Which data set has the highest number of correct predictions under a logistic regression model?
   2. Which data set has the lowest number of correct predictions?
   3. For which data points did *all three data sets* lead to *incorrect* predictions?
   4. Create the following table for each data set that you modeled using logistic regression (3 tables in total). Each cell must contain the number of predictions that fall into that category. For example, the cell on the top left is a count of the number of instances when the bank note is in reality forged *and* the model also predicts that it is forged. The cell on the bottom left is a count of the number of instances when the bank note is in reality genuine *but* the model predicts it is forged. And so on for the other cells.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted Value | |
|  |  | Forged (0) | Genuine (1) |
| Actual Value | Forged (0) |  |  |
| Genuine (1) |  |  |