**Homework Part 2**

**Linear and Non-Linear 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 regression on the bikeshare-daily.xlsx dataset.

Specifically, carry out the following steps:

1. Use the feature named “total” as the target feature. This is the output that you’re interested in predicting using regression with multiple features and then non-linear regression. (Note: This is just a step to take and there are no questions to answer)
2. Use Orange’s File widget to designate each feature as categorical or numeric. Make “instant” a meta feature. Skip the “casual” and “registered” features because they add up to the “total”. (Note: This is just a step to take and there are no questions to answer)
3. Visualize the features using a scatter plot in a systematic way. Select categorical-categorical, categorical-numeric, and numeric-numeric comparisons. In particular, create 6 scatter plots using your choices of features to display and write a sentence or two describing each scatter plot. Make sure you capture the images and paste them into your Word document.
4. Create a linear regression model using all the features you’ve selected using the default regression settings (“No regularization”). (Note: This is just a step to take and there are no questions to answer)
5. What are the optimal values of the parameters? In Orange w0 is called the “intercept”. The names of the other coefficients will be the same as the feature names. Report these values in the Word document and save them values in an Excel file. (Use Orange’s Save Data widget.)
6. Use the model and data to make predictions on your data. In other words, for each set of feature values in the dataset, find the predicted total number of bikes rented. Compare the predicted values with the original totals in the dataset using a scatter plot. Save these values in an Excel file.
7. Now simplify the dataset and create a dataset that only includes the features “atemp”, “hum”, “windspeed”, and “total”. Use “total” as the target feature. (Note: This is just a step to take and there are no questions to answer).
8. Split this simplified dataset – put 80% into a dataset we’ll call the “training” dataset and 20% into a dataset we’ll call the “test” dataset. Use Orange’s Data Sampler widget to do the 80-20 split. (Note: This is just a step to take and there are no questions to answer).
9. For the *training dataset*, create the following regression models (use the default “No regularization” setting): (Note: This is just a step to take and there are no questions to answer)
   1. y = (w1 \* atemp) + (w2 \* hum) + (w3 \* windspeed)
   2. y = (w1 \* atemp^2) + (w2 \* hum/windspeed) – (w3 \* windspeed^3)
   3. y = (w1 \* atemp \* windspeed) + (w2 \* hum \* windspeed^(1/2)) – (w3 \* windspeed^5)
10. For each model, capture the coefficients of the model in an Excel file. Use these coefficients to make predictions on the *test dataset*. You should end up with 3 Excel files – one each for predictions on the test dataset using models 9a, 9b, and 9c above.
11. Compare the predictions of the models using side-by-side scatter plots of the actual versus predicted values in the dataset. How do the non-linear models compare when it comes to prediction? Describe your results.
12. Suppose you are the head of business strategy for the bike share company. List 3 business questions you would use this dataset (the entire original dataset you started with) to answer.