Lab 10 Report

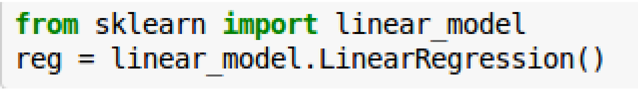
The aim of this lab is to predict the correct flow measurement based on the same datasets of Lab 9 as well as the probability (density) that the corresponding measurement is correct. The methodology is ensemble modeling using 3 separate methods and merge their results for final output. This report discusses the entire processing of the datasets and presents our results and findings in a brief manner. Same as last lab, the report is structured in three aspects:

1. Observations, intuitions, ideas

The datasets are the same with Lab 9 with parameters of flow, speed and occupancy for 5 different zones. Plus, the output of Lab 9, which is the probability density of each measurement combination. In this lab, we are supposed to predict correct flow values only using ensemble model, with 3 different methods. The general idea is to use the 3 methods to predict the value respectively along with the confidence of the value, and then merge the 3 results to a final result using confidence as the weight. Same as Lab 9, there are some empty or negative values in the datasets which we need to preprocess before fit in a model.

2. Analytics, statistical approaches and implementation

The method is given in Lab 10 instruction. In method 1, we fit a linear regression model for flow data of the lane that we want to predict and the lane next to it. We used:



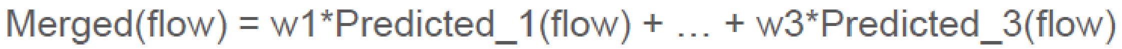
```
from sklearn import linear_model
reg = linear_model.LinearRegression()
```

to train the model and then predict the new values backwards using the model. Two details are described as follows. First, before fitting the model, we removed all the rows with empty and negative values in the datasets in order to avoid noisy data interference. Second, we only used one lane of data that is on the right side of the lane we want to predict for model fitting. For the rightmost lane, we just used the first lane for model fitting. Specifically for zone 3451, there is only one lane of data in the file, so we ignored the results of method 1 as there is no nearby lanes to build the model.

In method 2, we followed the rules in the instruction, which used the data of rows above and below the row we want to predict for value calculation. For the first row which has no row above, we just used the second row for predicting purpose. Similarly for the last row we used the second to the last row for prediction.

Method 3 is straightforward, which only used original data as part of the contributions to the final results.

In the final merging step, the given formula is used for final output calculation, which is



$$\mathrm{Merged(flow)} = \mathrm{w1*Predicted\_1(flow)} + \ldots + \mathrm{w3*Predicted\_3(flow)}$$

While w1, w2, w3 is the weight presented by confidence values of each method.

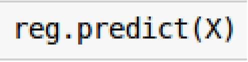
3. Key results, findings and challenges

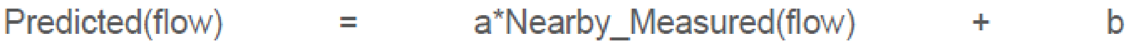
Some of the key results while running the program are as follows:

1). When cleaning the raw data, we found large volume of "bad" data. For example, for zone 1160, there are 3775351 rows of data originally while the non-negative usable data is 3663593 rows.

2). The RMSE (RMSE_FINAL) of final merged results and actual values is generally smaller than the RMSE (RMSE_N) of any method result and actual values. For example, for zone 1160 lane 1, RMSE_FINAL is 9.32, while the RMSE_1 is 9.46. This shows the ensemble model is working better than any single methodology.

From the final output, we could tell that the numbers listed are more reasonable and closely distributed. Also,

In this lab, we also met some challenges. First, we have to generalize the program so that it is usable for each different zones as they all have different numbers of lanes. Second, as we built the linear regression model in method 1, when we want to use  `reg.predict(X)` to directly get predicted values, sometimes it did not work since there are some "NAN" values in some zones. Therefore we manually calculate the predicted value using:



$$\mathrm{Predicted(flow)} \quad = \quad \mathrm{a*Nearby\_Measured(flow)} \quad + \quad \mathrm{b}$$

where a and b could be obtained directly from the model. Then we leave the NAN rows to the original values.