EP2300: Management of Networks and Networked Systems

Estimating Conformance to Service Level Agreements (SLAs) using Machine Learning

Task II - Estimating Service Metrics from Device Statistics

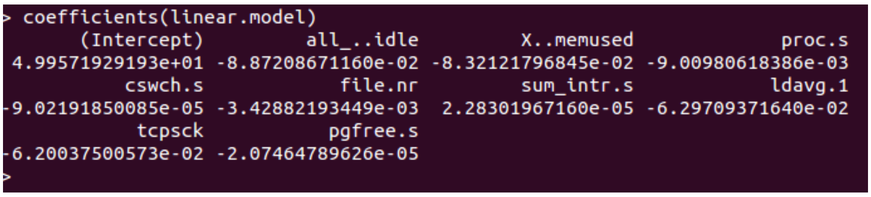
Surya Seetharaman

September 26, 2016

1. **Evaluate the Accuracy of Service Metric Estimation**

   a) Model Training - use linear regression to train a model M with the training set. Provide the coefficients (θ_1 , ..., θ_9) of your model M .

   Ans : First we use the sample() function in R to divide the given observations into training set and test set. Then using lm() and predict() functions, we can train the model through linear regression. The coefficients of the model are :



```
> coefficients(linear.model)
      (Intercept)         all_..idle         X..memused             proc.s
 4.99571929193e+01 -8.87208671160e-02 -8.32121796845e-02 -9.00980618386e-03
          cswch.s            file.nr         sum_intr.s            ldavg.1
-9.02191850085e-05 -3.42882193449e-03  2.28301967160e-05 -6.29709371640e-02
           tcpsck           pgfree.s
-6.20037500573e-02 -2.07464789626e-05
>
```

Figure 1: Output values of the Coefficients of the model

   b) Accuracy of Model M - compute the estimation error of M over the test P set. We define the estimation error as the Normalized Mean Absolute Error (NMAE) $= 1/\bar{y}(1/m \sum_{i=1}^{m} |y_i - \hat{y}_i|)$ whereby, $\hat{y}_i$ is the model estimation for the measured service metric y_i , and $\bar{y}$ is the average of the observations y_i of the test set, which is of size m = 1080. We consider an estimation accurate if NMAE < 15%.

   Ans : The estimation error (NMAE) was found to be **0.10152934116** which is approximately 10% and as it is less than 15%, we can consider this estimation to be accurate.

   c) Produce a time series plot that shows the measurements and the model estimations for the Video Frame Rate values in the test set.

Ans : This plot shown in Figure 2 was produced using the ordinary plot() function in R. We can see that the measured values (red) and estimated values (blue) almost coincide or show approximately the same kind of pattern though the measured rates are more distributed then the estimated values which are more denser.

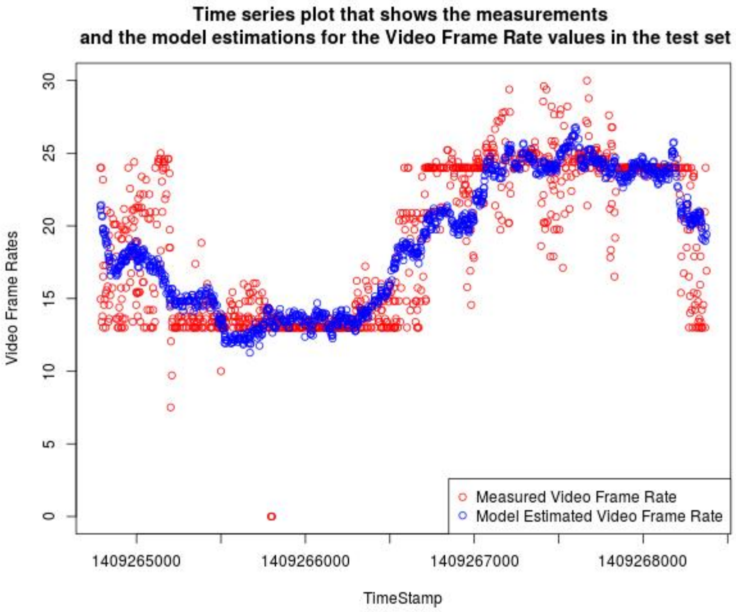


Figure 2: Time Series plot portraying the measurement and model estimation values for Video Frame Rate values in the test set.

d) Produce a density plot for the Video Frame Rate values in the test set.

Ans : This plot shown in Figure 3 has been produced using the density function in R and here we can see that there are two high peaks which mark the most dense areas of the video frame rates.

e) Produce a density plot for the NMAE values in the test set.

Ans : The density plot for the absolute errors is shown in Figure 4 and here you can see that the density peak is right at the beginning of the plot.

f) Discuss the accuracy of estimating the Video Frame Rate in this way.

Ans : So the accuracy of estimating the video frame rates using linear regression when calculated, the NMAE came to around 10%. Since it was less than 15% (which is the limit specified in this scenario) it is considered to be accurate. But still the error rate is 10% which cannot be ignored and it is also because our model is trained on a smaller data set (after excluding 30% of test data) and its error is likely to be higher than if we trained it on the full data set. But there are also other non-linear methods since there are times when linear regression cannot perform a good fit for the measured values. So this in scenario, where the NMAE value should only be < 15% for it to be accurate and our NMAE is 10%, we can say that this method of estimating the video frame rate values in accurate here.

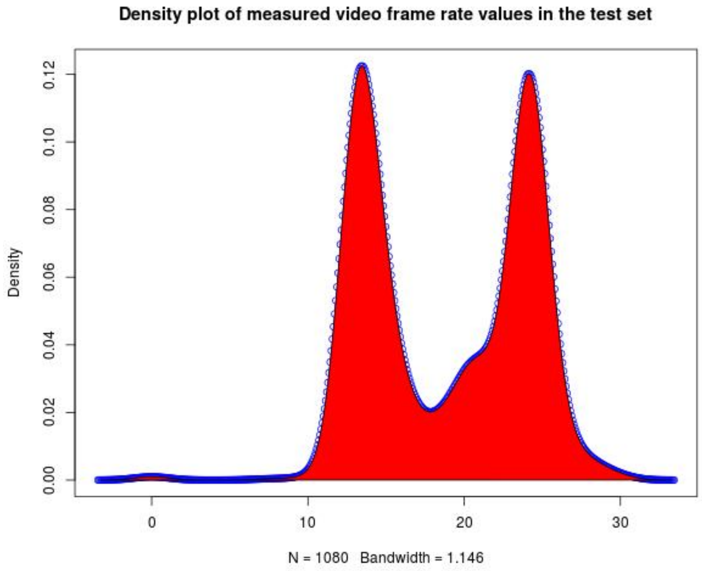


Figure 3: Density plot for measured Video Frame Values in the test set.

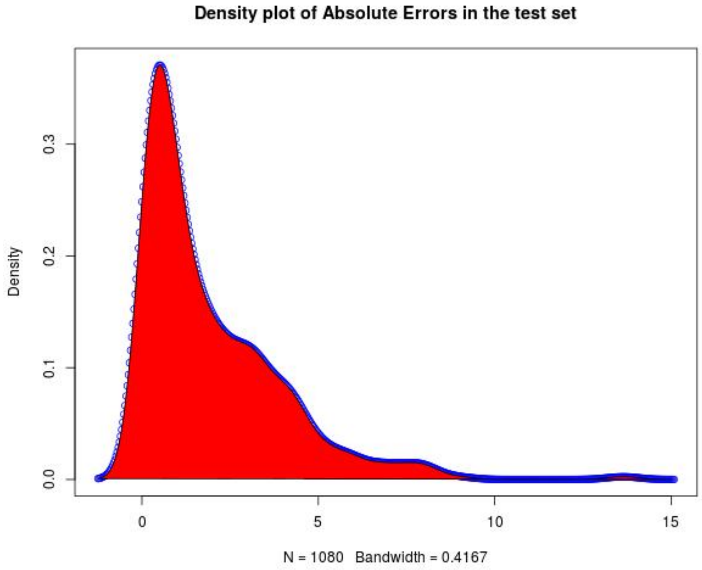


Figure 4: Density plot for absolute difference in the test set.

2. **Study the Relationship between Estimation Accuracy, the Size of the Training Set, and the Model Training Time.**

   a) Model Training - use the same linear regression method as above to train five models M_1 , ..., M_5 , one for each training set. Provide a table with the coefficients of these models.

   Ans : The coefficient values for all the five models was found using the coefficient() function in R and it is given in Table 1.

Table 1: Coefficient values for the five models trained using linear regression

Title	model50	model500	model1000	model1500	model2520
Intercept	7.61414500110e+01	4.31889405628e+01	4.83893873912e+01	5.13888426809e+01	4.99571929193e+01
all_..idle	-1.32239472395e-01	6.88663455684e-02	-9.87784587914e-02	-8.45331626590e-02	-8.87208671160e-02
X..memused	-7.43233463458e-02	-4.87084183257e-02	-6.04562331058e-02	-7.43337523959e-02	-8.32121796845e-02
proc.s	2.01216320822e-02	-3.22543350506e-02	-1.58547014476e-02	-1.62769985849e-02	-9.00980618386e-03
cswch.s	-2.85357849669e-04	-4.25292869336e-05	-9.25750833146e-05	-8.75011607939e-05	-9.02191850085e-05
file.nr	-1.19647515563e-02	-2.76748263446e-03	-3.49276608035e-03	-4.52846832834e-03	-3.42882193449e-03
sum_intr.s	-3.90603013125e-05	5.94246196496e-05	-4.39152862747e-06	8.53628631450e-06	2.28301967160e-05
ldavg.1	3.78286901164e-02	-7.70354599948e-02	-6.07767898526e-02	-6.07472499501e-02	-6.29709371641e-02
tcpsck	-8.89873150975e-02	-6.24952672522e-02	-6.26391465398e-02	-5.18661566736e-02	-6.20037500573e-02
pgfree.s	-2.80365543551e-06	-2.38696352690e-05	-1.51428371153e-05	-1.86613494188e-05	-2.07464789626e-05

   b) Training Time of the Models - on your computer, measure the execution time (in milliseconds) to train each of the five models.

   Ans : The execution time for training the five models, i.e using linear regression (lm() function execution time) was measured using the system.time() function in R. The values (in milliseconds) are given in Table 2

Table 2: Execution time for training the five models in my system (in milliseconds)

Title	model50	model500	model1000	model1500	model2520
user	0	4	4	4	8
system	0	0	0	0	0
elapsed	**3**	**4**	**5**	**6**	**7**

   c) Accuracy of Models - compute the NMAE for each of the five models for the test set.

   Ans : The NMAE values for the five models are as shown in Table 3.

   d) Produce a plot that shows both the NMAE and Training Time for all models. Use two y-axes, one on the left side and one on the right side, to include both curves in a single plot.

Table 3: NMAE values of the five models

Title	model50	model500	model1000	model1500	model2520
NMAE	0.116184750047	0.101951827430	0.102588853654	0.101616573353	0.101529341160

Ans : The plot is shown in Figure 5. The left side is the y axis for NMAE while the right side y axis marks the values of training time.

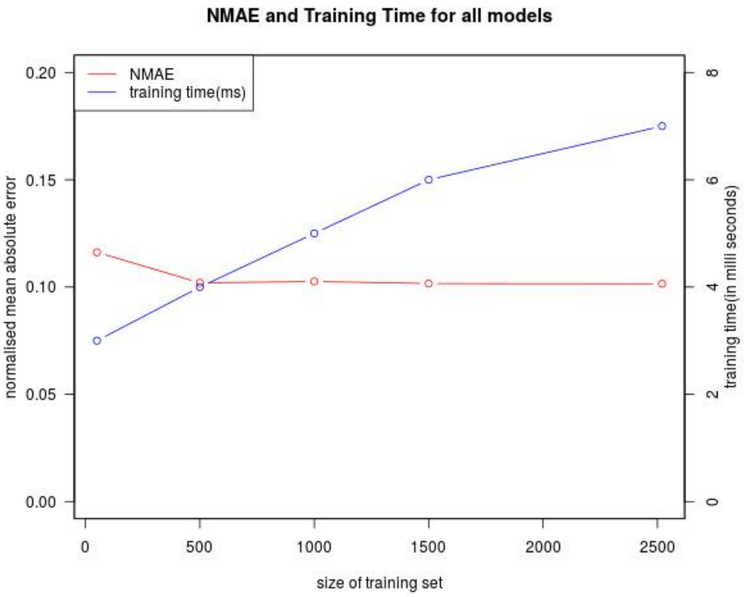


Figure 5: Plot that shows both the NMAE and Training Time for all models

e) Discuss the relationship between the number of observations in the training set and the accuracy of the model estimations.

Ans : As the number of observations in the training set increases, the accuracy also increases, i.e the error rate is less as can also be seen from the plot given in Figure 5. This is because as the size of the training set increases, there are more values to look at and come to an estimation due to which there is less chances of error.

f) Discuss the relationship between the number of observations in the training set and the training time of a model.

Ans : As the number of observations increases, the training time also increases since it takes more time now to go through all the measured values in the training set and then compare and train the model. So larger the size of the training set, more time it takes to train it.

Note : The program can be run using the command **R --vanilla < task2.R --args X.csv Y.csv** , task2.R being the name of the source code file.