EP2300: Management of Networks and Networked Systems

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

Task III - Estimating SLA Conformance and Violation from Device Statistics

Surya Seetharaman

October 8, 2016

1. **Evaluate the Accuracy of SLA Estimation, Training time, and their dependence on Training Set Size**

   a) Classifier Training - use logistic regression to train five classifiers C1 , ..., C5 , one for each training set. Make a table showing the coefficients (θ) of the classifiers;

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

Table 1: Coefficient values for the five classifiers trained using logistic regression

Title	C_1(50)	C_2(500)	C_3(1000)	C_4(1500)	C_5(2520)
Intercept	-1.53391992521e+03	2.03842085790e+01	3.01747389961e+01	2.27959114161e+01	2.24966385670e+01
all_..idle	-1.36541782466e+01	-1.05721647591e-01	-1.15596200227e-01	-1.15866772756e-01	-1.07038994389e-01
X..memused	2.37981112866e+00	-1.04719255296e-01	-1.25510043651e-01	-9.87112680648e-02	-9.55740242066e-02
proc.s	-3.75264688597e+00	-3.58083760378e-03	-9.97893720041e-03	-9.13779139306e-04	-4.23648517052e-03
cswch.s	-1.11797949240e-02	-5.90600975298e-05	-8.63582119157e-05	-8.63068159437e-05	-7.02670650111e-05
file.nr	5.87723003076e-01	5.87723003076e-01	-1.42450837901e-03	-3.22799547092e-03	-9.65010123972e-04
sum_intr.s	8.12399969903e-02	3.18234073364e-04	1.96328583972e-04	1.38713626347e-04	1.56841055692e-04
ldavg.1	-1.12884234769e+01	-6.44686810520e-02	-5.99037748710e-02	-5.91762918476e-02	-6.26286585649e-02
tcpsck	3.95137794235e+00	-5.65525359104e-02	-5.06756709415e-02	-5.21257447812e-02	-5.47781607203e-02
pgfree.s	-6.74565483804e-04	-1.37262638261e-05	-1.38438434508e-05	-1.37674765532e-05	-1.47038163325e-05

   Note : When using glm(); only for the classifier with training set size fifty, I got a warning message sometimes, no matter whichever seed I used (I did try using several seeds as mentioned in Bilda under discussions) .

   b) Time to Train the Classifiers C_1 , ..., C_5 - measure the execution time in milliseconds to train each classifier;

   Ans : The execution time for training the five models, i.e using logistic regression (glm() 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	C_1(50)	C_2(500)	C_3(model1000)	C_4(1500)	C_5(2520)
user	4	8	16	24	36
system	0	0	0	0	0
elapsed	**6**	**10**	**16**	**24**	**37**

c) Accuracy of the Classifiers C 1 , ..., C 5 - Compute the classification error (ERR) on the test set for each classifier. Consider a classifier as accurate when ERR $<$ 15%.

Ans : The ERR values for the five classifiers are as shown in Table 3.

Table 3: ERR values of the five models

Title	C_1(50)	C_2(500)	C_3(1000)	C_4(1500)	C_5(2520)
ERR	0.136111111111	0.103703703704	0.105555555556	0.104629629630	0.102777777778

d) Produce a plot that shows both the Classification Error (ERR) on the test set and Training Time for each of the five classifiers C_1 , ..., C_5 . Use two y-axes to include both curves in a single plot.

Ans : The plot is shown in Figure 1. The left side is the y axis for ERR while the right side y axis marks the values of training time.

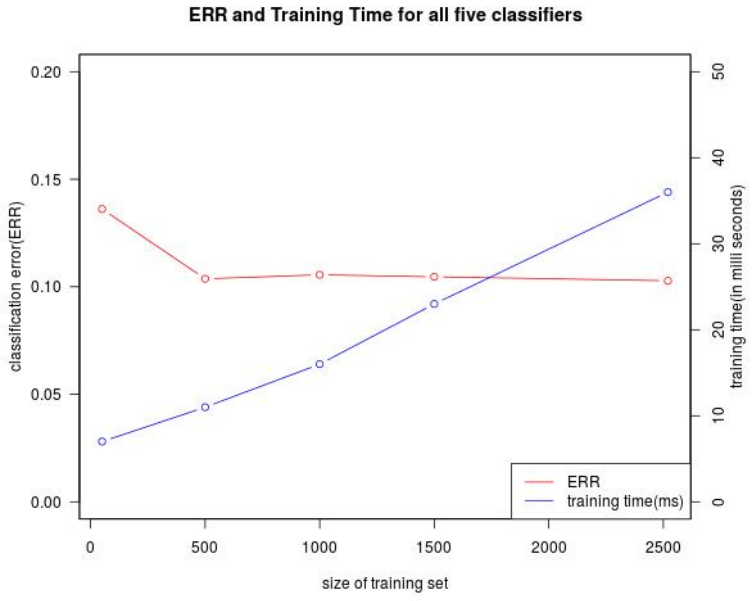

Figure 1: Plot that shows both the ERR and Training Time for all classifiers

e) Using the test set produce a time series plot that shows on the y-axis the observed video frame rate. For each point in the plot indicate the classification obtained with the most

accurate classifier C_i . Use different symbols (or colors) for a correctly classified and a incorrectly classified observation.

Ans : The plot is shown in Figure 2.

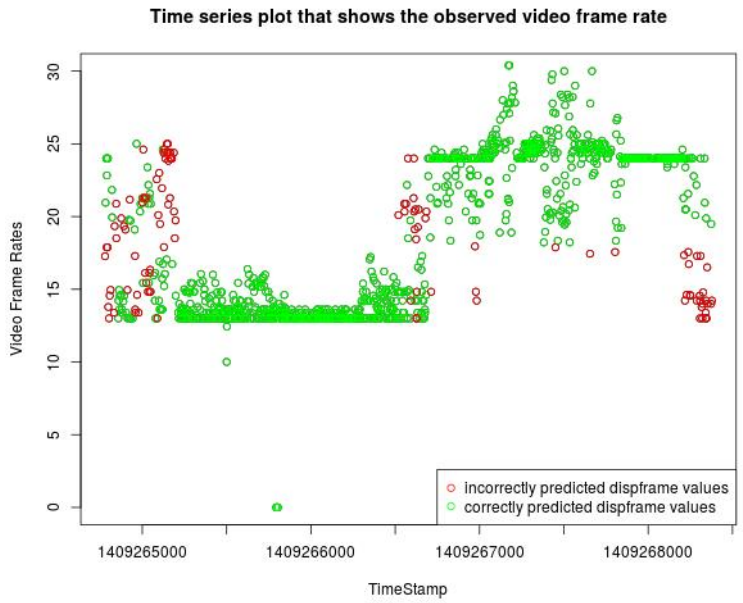

Figure 2: Time series plot that shows the observed video frame rates

f) Using the above plot discuss how the estimation accuracy relates to the evolution of the video frame rate over time.

Ans : Looking at the graph, the red points are the incorrectly predicted video frame rates while the green ones are the correctly predicted ones. There are some errors concentrated at the beginning, at the very end and some right in middle region. There are very few falsified rates, while most of them are green. On seeing the pattern, we can see that the falsified rates are occurring around the region of video frame rate = 18. This is the border where the rates are classified as having conformed to SLA or violated SLA. Thus there is room for more errors in this region. So estimation accuracy is less and errors are more around this region of video frame rate value being 18 over time. Also we can see that the red points are not clustered like green ones, which shows that with change in video frame rates, there is also a tendency for errors i.e less accuracy. So basically rapid changes and border value 18 are difficult scenarios for classification due to which it is more prone to errors in these situations throughout time. Also maximum errors are in the beginning and with time they seem to be decreasing.

g) Discuss the relationship between the number of observations in a training set and the estimation accuracy for C_1 , ..., C_5.

Ans : As the number of observations in the training set increases, the accuracy of the classifiers also increases usually. 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. But as seen in the Figure 1, due to randomness in selection of training sets, there could also be cases where the accuracy is better in case of smaller sized training sets than the larger sized training sets though with not much of a difference. So for logistic regression case, here it might be difficult to judge.

h) Discuss the relationship between the number of observations in a training set and the training time for C_1 , ..., C_5 .

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 as is seen from Figure 1.

Note : The program can be run using the command **R --vanilla < task3.R --args X.csv Y.csv** , task3.R being the name of the source code file.

EP2300: Management of Networks and Networked Systems

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

Task IV - Comparing the Accuracy of two methods for SLA Estimation

Surya Seetharaman

October 11, 2016

1. **Accuracy of SLA Estimation: C_{lin} versus C_{log}**

   a) Classifier Training - train five classifiers $C_{lin1}, ..., C_{lin5}$, one for each training set. Do the same for the classifiers $C_{log1}, ..., C_{log5}$. Draw a table showing the coefficients (θ) of all ten classifiers.

   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 models through linear regression after which based on the results, we assign the binary values 0 or 1. The coefficients of the models are :

Table 1: Coefficient values for the five linear classifier models.

Title	C_{lin1}(50)	C_{lin2}(500)	C_{lin3}(1000)	C_{lin4}(1500)	C_{lin5}(2520)
Intercept	3.39501464447e+01	5.40898625952e+01	4.89731692834e+01	-9.97551157753e-02	4.87974318901e+01
all_..idle	-8.87430817771e-02	-9.89586260893e-02	-9.97551157753e-02	-9.89307612239e-02	-8.46818753300e-02
X..memused	-6.62873826500e-03	-1.00383174212e-01	-1.18097895060e-01	-1.13104873323e-01	-9.20572678862e-02
proc.s	-1.92686715258e-02	-9.25464074633e-03	-6.65639279443e-03	1.08890496563e-02	-2.25750051624e-03
cswch.s	-7.66599424128e-06	-1.14643609877e-04	-8.26317889697e-05	-1.06160818222e-04	-7.89740768763e-05
file.nr	-7.00709034467e-04	-4.10060199743e-03	-1.69107823947e-03	-3.41278181390e-03	-3.02924090065e-03
sum_intr.s	7.01318371078e-05	1.10937290740e-05	4.30449084655e-05	2.19953648593e-05	4.60229043507e-05
ldavg.1	-9.33738123495e-02	-5.44465425505e-02	-7.20013141132e-02	-5.94736778229e-02	-6.47560490883e-02
tcpsck	-5.45516465693e-02	-5.92455268537e-02	-7.25268984093e-02	-6.69342148288e-02	-6.57632754448e-02
pgfree.s	-3.67217141025e-05	-1.94300606577e-05	-2.21239058050e-05	-1.93459615146e-05	-2.01709799977e-05

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

Table 2: Coefficient values for the five classifiers trained using logistic regression

Title	C_1(50)	C_2(500)	C_3(1000)	C_4(1500)	C_5(2520)
Intercept	-7.12843017421e+01	2.71049412567e+01	2.77490020953e+01	2.67900999973e+01	2.32363550795e+01
all_..idle	2.61325485383e+02	-1.38287139202e-01	-1.35616396612e-01	-1.04753543122e-01	-9.17007824939e-02
X..memused	1.37499389074e-01	-1.16416989630e-01	-1.33272243610e-01	-1.21955734840e-01	-1.07344548157e-01
proc.s	7.86104559571e-02	6.49738341547e-04	-9.86927102774e-03	1.39879893265e-02	-1.33692315714e-03
cswch.s	2.59635902893e-04	-1.09391406794e-04	-1.05048351662e-04	-8.21315368092e-05	-5.37540071066e-05
file.nr	7.57648550742e-03	-2.13662750836e-03	-7.95288366815e-04	-2.10439293650e-03	-1.63632312540e-03
sum_intr.s	4.08237099595e-03	2.68375818445e-04	7.30046545245e-05	1.53449125517e-04	1.63824364276e-04
ldavg.1	-1.12884234769e+01	-5.95376743277e-02	-6.87647755789e-02	-6.35772037603e-02	-6.66155092234e-02
tcpsck	-3.34502026653e-01	-4.44497503934e-02	-3.65681045347e-02	-3.82134112475e-02	-1.26261013106e-05
pgfree.s	-4.78810092192e-05	-2.28604221277e-05	-1.43522109768e-05	-1.26261013106e-05	-1.53996340542e-05

b) Accuracy of the Classifiers - Compute the classification error (ERR) for C_{lin1} , ..., C_{lin5} and C_{log1} , ..., C_{log5} . Consider a classifier as accurate when ERR < 15%.

Table 3: ERR values of the five C_{log} models

Title	C_{lin1}(50)	C_{lin2}(500)	C_{lin3}(1000)	C_{lin4}(1500)	C_{lin5}(2520)
NMAE	0.1064814814815	0.0907407407407	0.1055555555556	0.0962962962963	0.0953703703704

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

Table 4: ERR values of the five C_{lin} models

Title	C_{log1}(50)	C_{log2}(500)	C_{log3}(1000)	C_{log4}(1500)	C_{log5}(2520)
ERR	0.1518518518519	0.1018518518519	0.1111111111111	0.0981481481481	0.0990740740741

c) Produce a plot that shows on the y-axis the classification error in the test set, and on the x-axis the size of the training set. The errors of the classifiers C_{lin1} , ..., C_{lin5} form a curve on this plot, the errors of the classifiers C_{log1} , ..., C_{log5} form a second curve;

Ans : This plot shown in Figure 1 was produced using the ordinary plot() function in R.

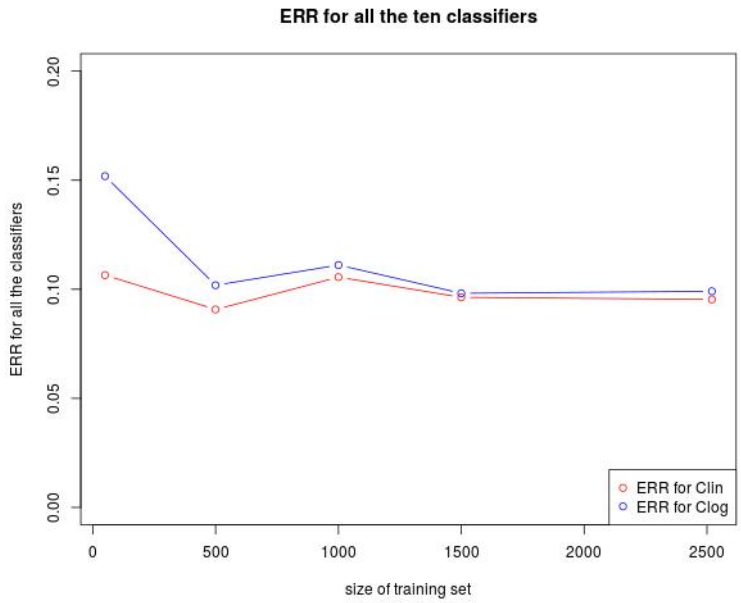

Figure 1: Classification errors in the test set for all the classifiers.

d) Using the above plot, compare the accuracy of the two types of classifiers C_{lin} and C_{log} . Can you explain the result?

Ans : Since for a classifier to be accurate, we have to have ERR below 15%. This is satisfied for all cases other than the C_{log1} as shown in Figure 1. In fact, both the types of classifiers show similar curves, in the sense that ERR decreases from 50 sized training set further for the 500 sized training set and then increases a little after which it decreases until the end. So actually with increase in size of training set, the accuracy should also increase and it is shown to get accurate in the plot with increase in size of training set, with the exception of it being more accurate at 500 sized set than the 2520 sized set, which could be due to the random sampling. So due to this unlikely behaviour, it is difficult to judge the behavior of accuracy here. As for the comparison between the two types, for both the types, the ERR is higher for 50 sized sample and then seems to get accurate towards 2520 sized sample, but both are most accurate the 500 sized sample. Also the curve for C_{lin} is lower than the curve for C_{log} and hence it has lesser ERR value.

Note : The program can be run using the command **R --vanilla < task4.R --args X.csv Y.csv** , task4.R being the name of the source code file.