Universidade Federal de Uberlândia

PGC308A Tópicos Especiais em Sistemas de Computação 2: Internet do Futuro

Rodrigo Moreira – moreira_r@outlook.com

2017-2

1 Task I: Data Exploration

1. Dataset statistics for each component of X and Y:

Table 1: Component Statitstics

Variables	Component	Mean	Maximum	Minimum	25th	95th	St. Dv.
X	all_..idle	9.06	70	0	0.0	38.62	16.12
	X..memused	89.14	98	73	82.97	96.77	8.18
	proc.s	7.68	48	0	0.0	20.0	8.53
	cswch.s	54045.87	83880	11398	31302.00	72135.10	19497.81
	file.nr	2656.33	2976	2304	2496.00	2880.00	196.11
	sum_intr.s	19978.04	35536	10393	16678.00	28228.40	4797.27
	ldavg.1	75.88	147	11	28.20	127.99	43.86
	tcpsck	49.00	87	21	34.00	71.00	15.87
	pgfree.s	72872.15	145874	15928	61601.75	97532.50	19504.32
Y	Video Frame Rate	18.82	30	0	13.39	24.61	5.22

2. Compute the following quantities:

   a – the number of observations with memory usage larger than 80% = **2875**

   b – the average number of used TCP sockets for observations with more than 18000 interrupts/sec = **46.35**

   c – the minimum memory utilization for observations with CPU idle time lower than 20% = **73.03**

3. Plots:

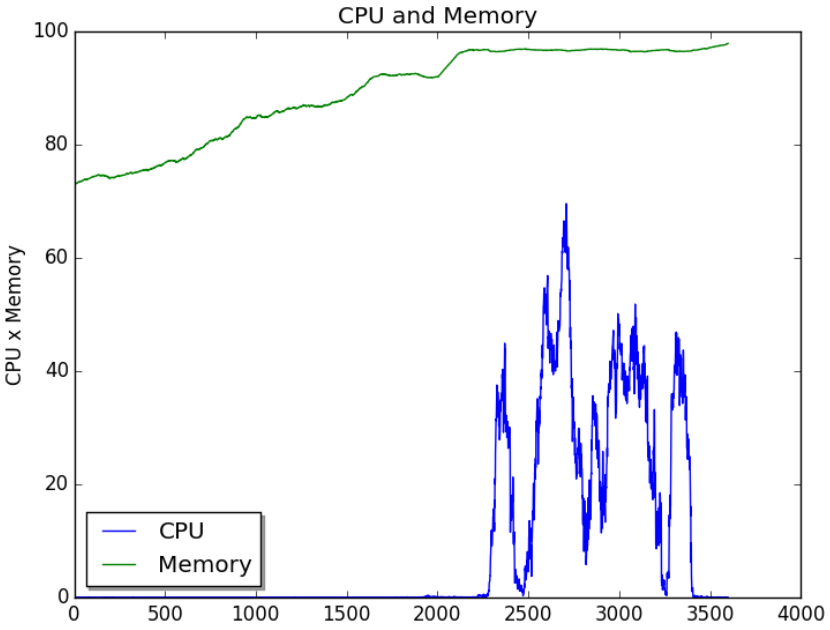

Figure 1: Time Series - CPU and Memory Idle

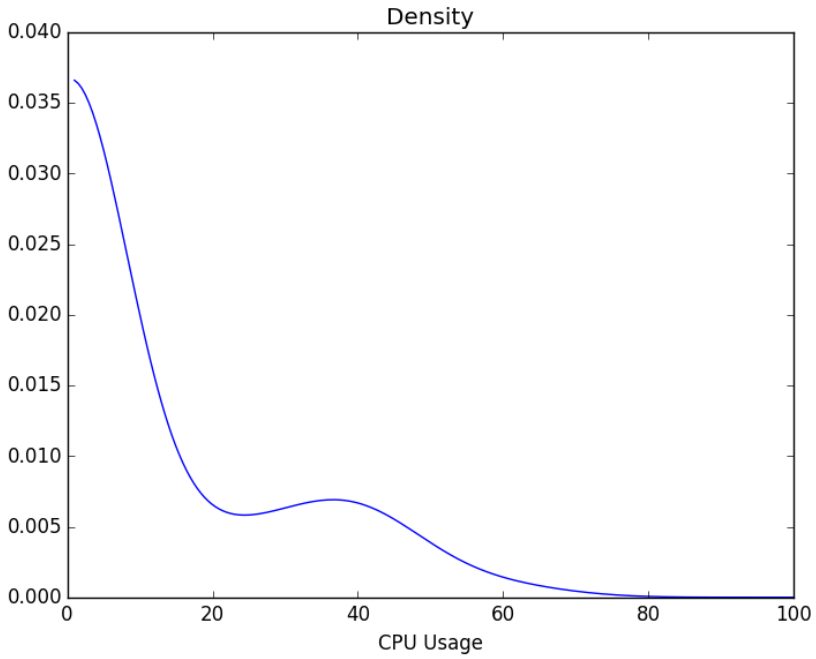

Figure 2: Density - CPU

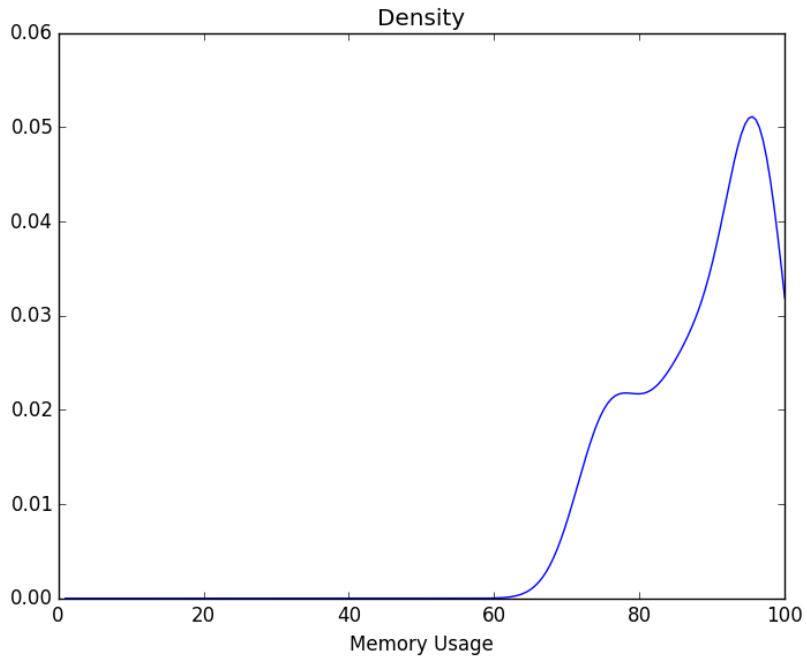

Figure 3: Density - Memory

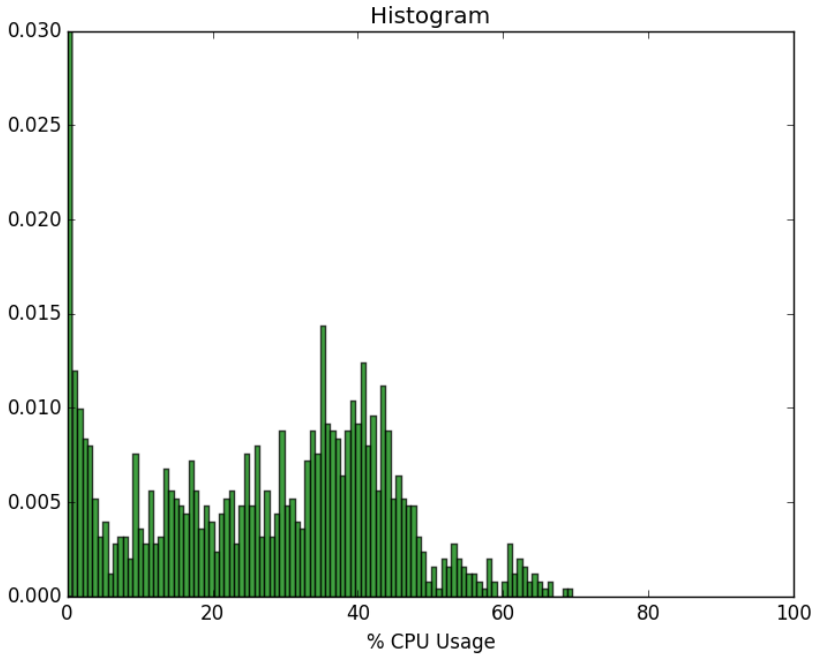

Figure 4: Histogram - % CPU

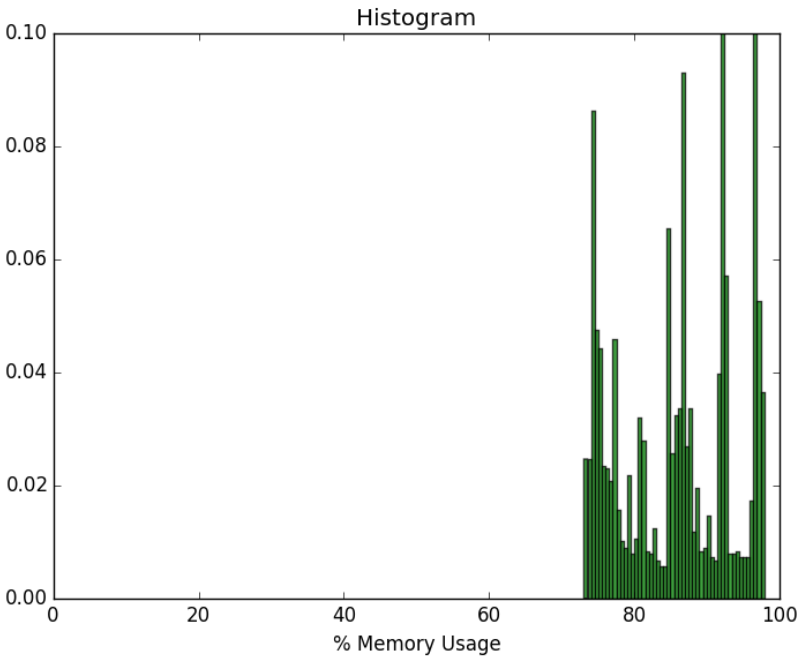

Figure 5: Histogram - % Memory

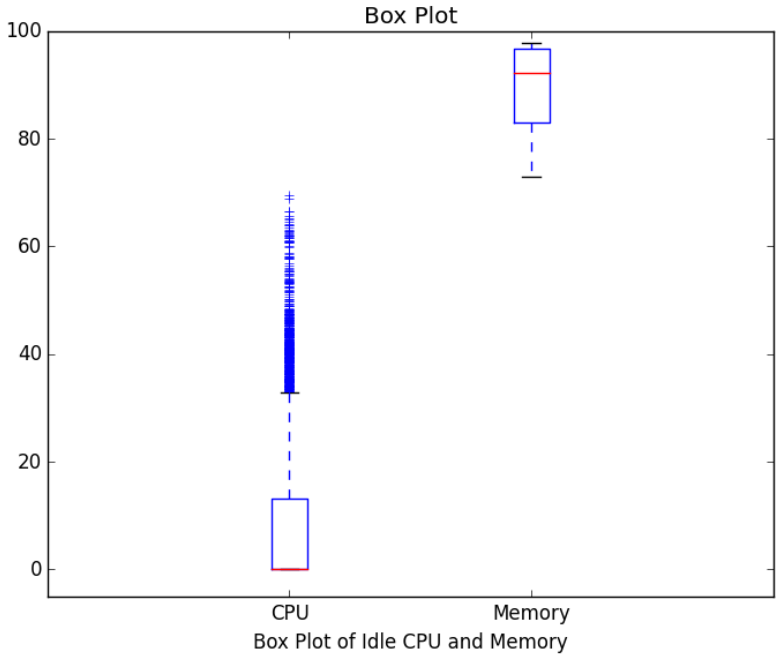

Figure 6: Boxplot - % CPU and Memory

2 Task II: Estimating Service Metrics from Device Statistics

1. Evaluate the Accuracy of Service Metric Estimation:

   a – Coefficients of the model M:
   Θ_1 = **-8.73835665e-02** Θ_2 = **-9.61953690e-02**
   Θ_3 = **-5.14847544e-03** Θ_4 = **–9.76028238e-05**
   Θ_5 = **-3.35999073e-03** Θ_6 = **2.88433117e-05**
   Θ_7 = **-6.01562113e-02** Θ_8 = **-6.34882988e-02**
   Θ_9 = **-2.07032275e-05**

   Note: Since this is a randomized evaluation, the values of the coefficients vary briefly from one execution to another.

   b – Accuracy of the model M:
   Regression Method: Normalized Mean Absolute Error is **0.0967**, that is 9.67%.
   Regression Method (Naïve-based): Normalized Mean Absolute Error is **0.2549**, that is 25.49%.

   c – The figure below depicts a BoxPlot of Measurements and the model Estimations. Also, the figure depicts a plot by considering the naïve method as baseline.

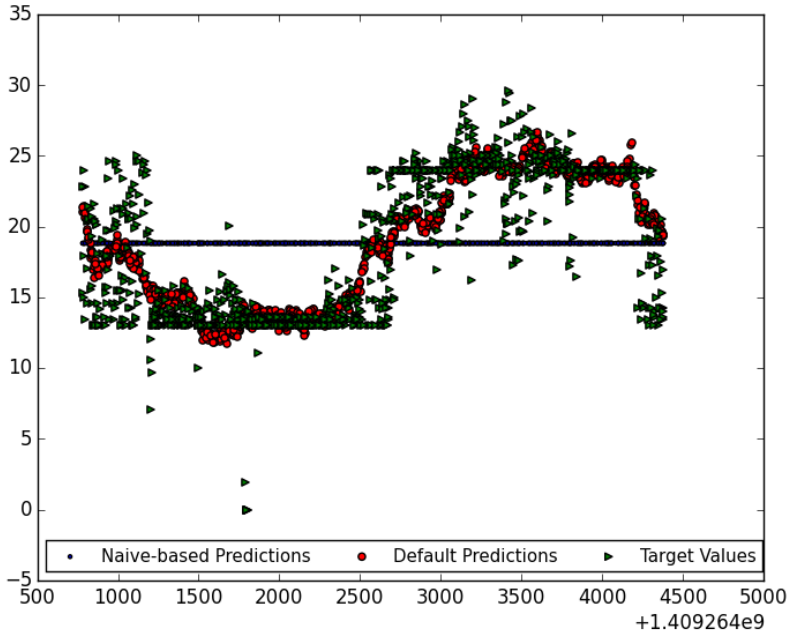

Figure 7: Default and Naive-based Regression

d – The figure below depicts a density plot and histogram for the Video Frame Rate values in the test by considering a bit size of histogram as 1 frame style:

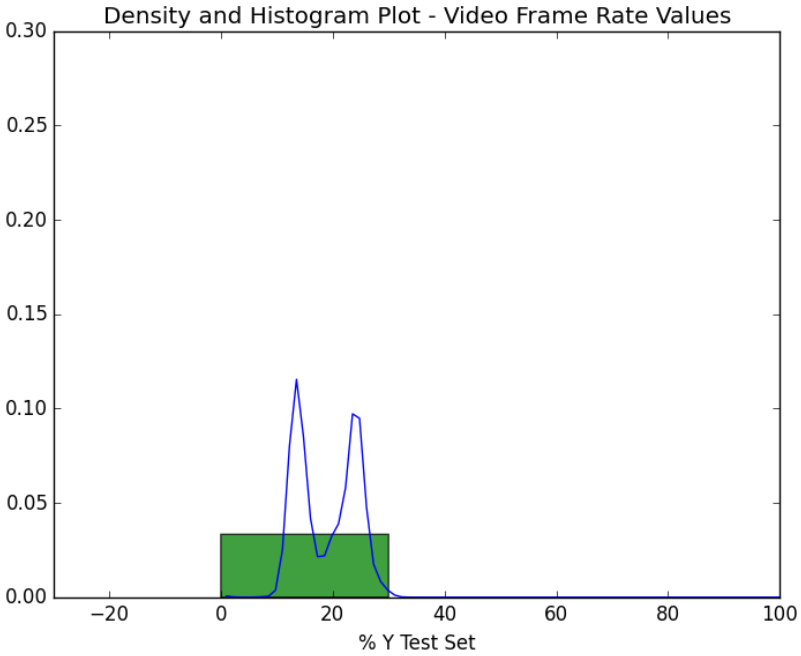

Figure 8: Histogram and Density - % Video Frame Rate

e – The next graph refers to the density of error in the test set.

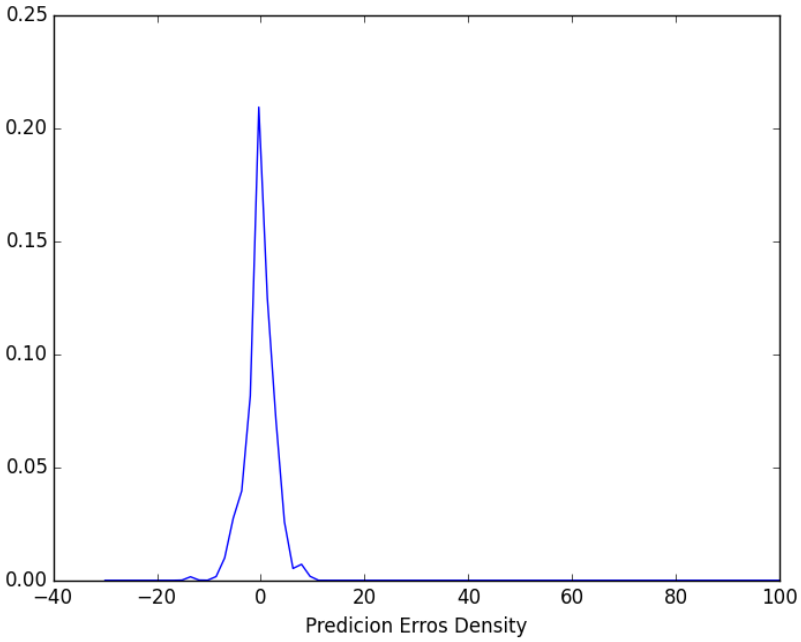

Figure 9: Density - % Errors of Prediction

f – We can observerve that for the video frame rate the prediction error is better in regression model than naïve method. The error measured by N.M.A.E is more smallest in training set that considers on the exact value of y_i than trainig set when the y_i is the same for all entry, in naïve method is the average of all y_i.

2. Study the Relationship between Estimation Accuracy and the Size of the Training Set

d –The next graph depicts tha NMAE for M against the size of the trainig set.

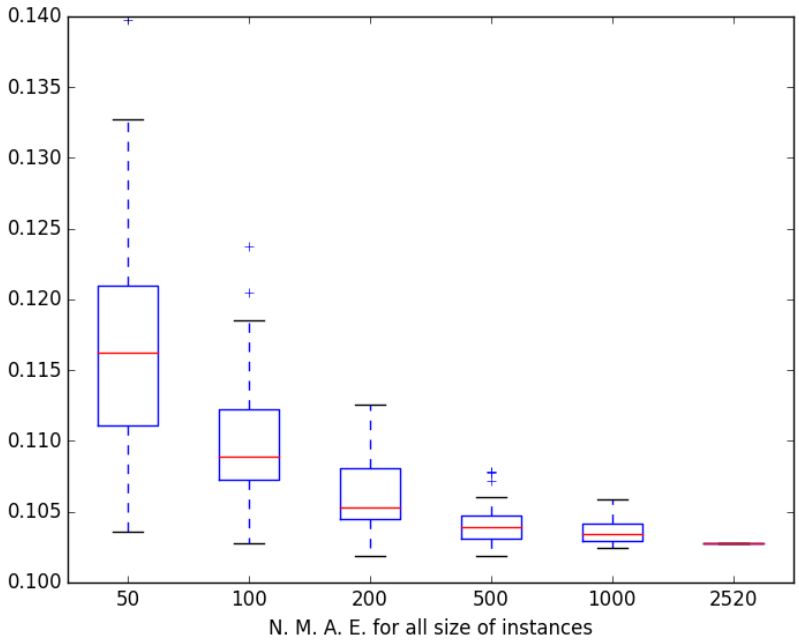

Figure 10: Boxplot - N.M.A.E against the size of training set

e – As we can see in the graph that after the size of training set increases the error measured through N.M.A.E decreases. This sugests that as we increases size of trainig set the accuracy was improved, but we do not formally measured yet, the model might is in overfitting case. To overcome this, it is necessary to run a cross validation.

The codes used to solve these questions are available in the following link: https://github.com/romoreira/MLN/