Weather Sensor Readings Predictions - Modeling & Inferencing in time series data

Shikha Chamoli

May 4, 2016

1 Problem Statement

Wireless sensor networks are increasingly being deployed to measure temperature, humidity, and light and to detect motion, fires, and landslides,among many other events of interest. Sensors are often operated on a battery and hence we often face a trade-off between acquiring frequent sensor readings versus maximizing their battery life. The biggest drainer of the battery is "communication." That is, sensors consume the most battery power not when they sense the environment but when they communicate their readings to a central server.

In this project, we train models of the sensor readings at a central server, and the central server will contact wireless sensors infrequently, to save battery. The central server will use actual readings of the sensors when it can obtain them, and it will use its own predictions when it can't obtain the readings.

2 Dataset

The dataset consists of temperature readings and humidity readings for 50 sensors for a total of five days. The dataset provided was split into train(3 days) and test(2 days).

3 Tasks

In this phase 3 this model is stationary at hour level. We do the forward inferencing as window inferencing and variance inferencing.

3.1 Modeling

We train a model of the sensor readings using training data and test it on the test data. In this model we consider that all the observations in train data (humidity/temparature) are independent.

3.2 Active Inference

Given a model of the sensor readings (trained on the train data), and given a budget determining what percentage of the readings can be obtained by the central server, determine which sensors' readings should be obtained to minimize overall error. We use two active inference approaches in this project.

3.2.1 Window

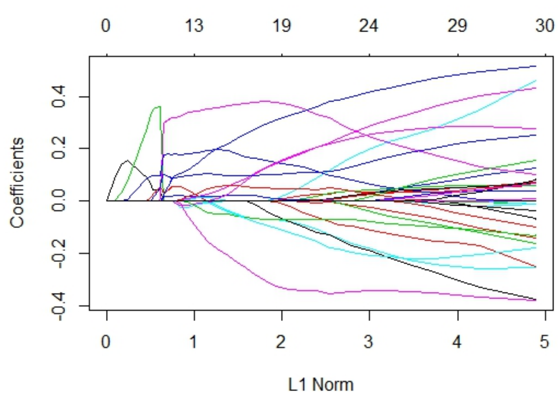
If the budget is 30%:

Time t: sensors 1-15
Time t+1: sensors 16-30
Time t+2: sensors 31-45
Time t+3: sensors 46-50 and 1-10

and so on will communicate their readings to the central server and central server will use its own predictions for the sensors that do not communicate their readings. Assume the central server has infinite memory so that it can remember all the readings it received in the past and current time.

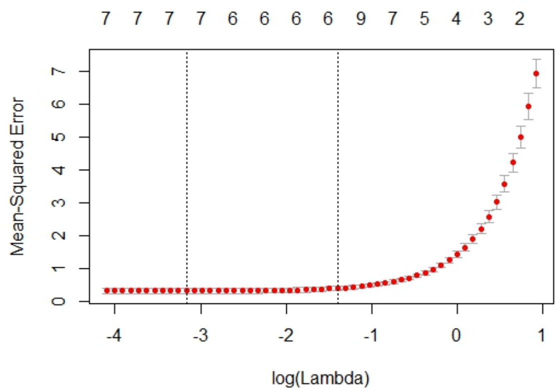
3.2.2 Variance:

If we assume the budget is 30%. At time t, the central server will chose 15 sensors for which it has the highest prediction variance. Follwing are the plots of beta parameters/ coefficients from this regression.



Graph 1- Lasso coefficients

Cross-validation curve (red dotted line) in the Graph 2 , and upper and lower standard deviation curves along the λ sequence (error bars). Two selected $\lambda's$ are indicated by the vertical dotted lines (see below).



Graph 2- Cross-validation curve (red dotted line)

3.2.3 Validating Lasso regression (given below by bb) against linear regression (olsReg) parameters

I checked the lasso regression params against linear regression params using **"tstat"**, and values were similar.

```
Console D:/IIT IITerm/cs583/MyProject/
Residuals:
     Min       1Q   Median       3Q      Max
-1.26584 -0.22183 -0.05802  0.13494  2.19456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.28714    0.81690   6.472 1.65e-09 ***
xxV2        -0.11079    0.13067  -0.848  0.39802
xxV7         0.38457    0.08448   4.552 1.17e-05 ***
xxV13        0.02444    0.06273   0.390  0.69749
xxV20        0.05304    0.04588   1.156  0.24969
xxV22        0.20650    0.05958   3.466  0.00071 ***
xxV40        0.08373    0.03657   2.289  0.02360 *
xxV48        0.09471    0.03395   2.790  0.00604 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5178 on 135 degrees of freedom
Multiple R-squared:  0.9628,    Adjusted R-squared:  0.9609
F-statistic: 499.6 on 7 and 135 DF,  p-value: < 2.2e-16

> olsReg$coefficients
(Intercept)        xxV2        xxV7       xxV13       xxV20       xxV22       xxV40       xxV48
 5.28714236 -0.11078931  0.38456764  0.02443612  0.05303880  0.20649597  0.08373457  0.09470525
> bb[bb>0]
[1] 6.302352364 0.010023807 0.321872829 0.006370842 0.042780854 0.169938567 0.096880406 0.042442242
>
```

4 Experiments and Analysis (Phase1, Phase2-Model1, Phase2-Model2, Phase3)

Following bar graphs have models from phase 1, model 1and model2 from phase2 , forward inferencing model Phase3 as X-axis and mean absolute error as Y-axis.

4.1 Lasso regression

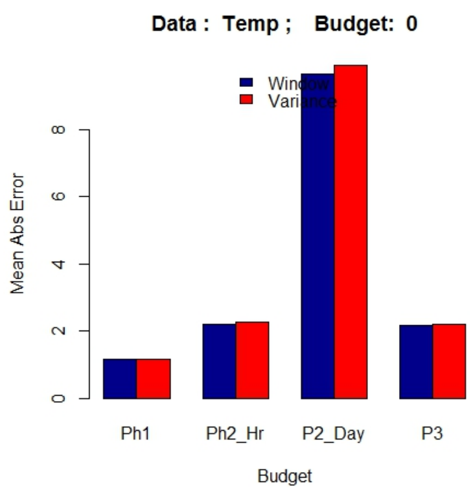
I used lasso regression here unlike in phase 1 and 2 where linear regression was used.

4.2 Temperature dataset plots for different budgets

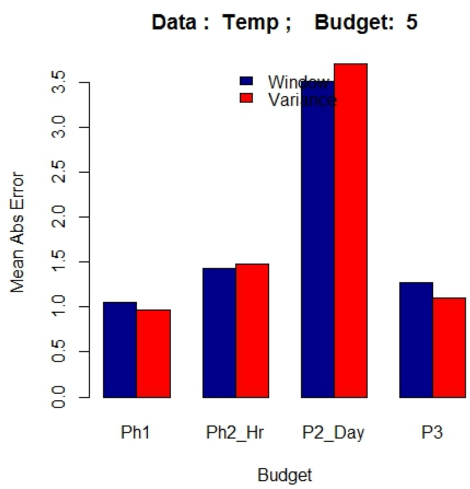
We take active inference budgets: [0, 5, 10, 20, 25]. These budgets indicate how many sensors at time t can communicate their readings to the central server. These are not percentages but are actual counts.

4.2.1 Notations

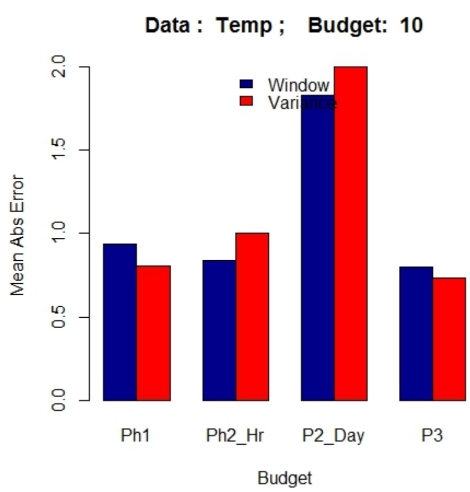
Ph1 for Phase1 ; Ph2_Hr for Phase2 model 1; Ph2_Day for Phase2 model 2; P3 for Phase3



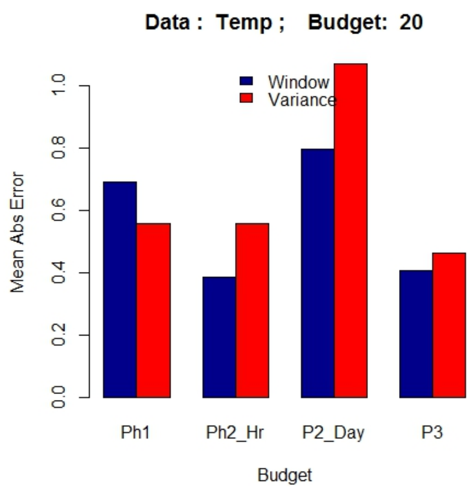
Plot 1- Phase1 & Phase2 models 1 & 2, Phase3 vs Mean Abs Error (Phase1, Phase2 Model2, Phase2 Model1, Phase3; Active inference (Window & Variance); Budget:0)



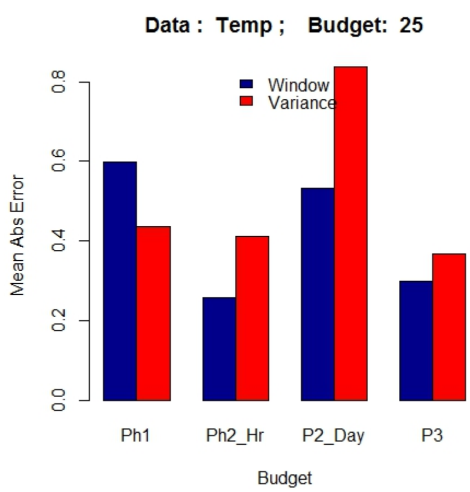
Plot 2- Phase1 & Phase2 models 1 & 2, Phase3 vs Mean Abs Error (Phase1, Phase2 Model2, Phase2 Model1, Phase3; Active inference (Window in Blue & Variance in Red); Budget: 5)



Plot 3- Phase1 & Phase2 models vs Mean Abs Error (Phase1, Phase2 Model2, Phase2 Model1; Active inference (Window & Variance); Budget:10)



Plot 4- Phase1 & Phase2 models, Phase3 vs Mean Abs Error (Phase1, Phase2 Model2, Phase2 Model1, Phase3; Active inference (Window & Variance); Budget:20)



Plot 5- Phase1 & Phase2 models, Phase3 vs Mean Abs Error (Phase1, Phase2 Model2, Phase2 Model1, Phase3; Active inference (Window & Variance); Budget:25)

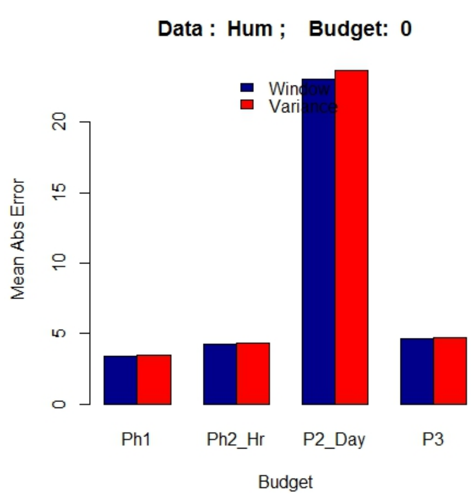
4.2.2 Comparison /Analysis (Temperature Dataset)

Looking at plots 1 to 5 for temeprature datasets we conclude the following:

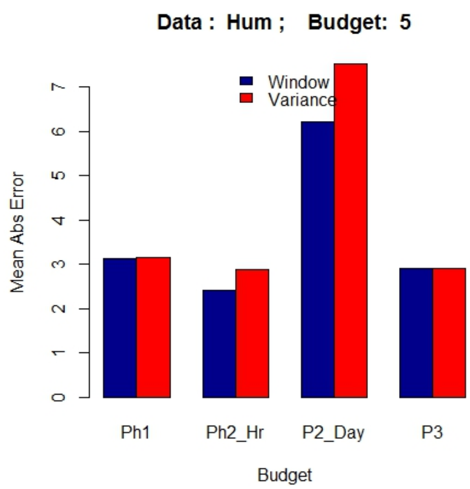
Phase 1	Phase2-Model 1	Phase2-Model 2
Performance of variance inferencing is better than window based.	Performance of window inferencing is better than variance based.	Performance of window inferencing is better than variance based.
With the increase in budget size absolute error decreases.	With the increase in budget size abs error decreases.	With the increase in budget size abs error decreases.
Best results seen from: this model, variance based method, when taken highest budget(25)	Best results is from window based method, when taken highest budget(25)	Best results seen: window based method, when taken highest budget(25)
(Yes) Phase1-Window method performs same as Phase 2-model2, Variance method for Budget=10	**Best performance out of all 4 models for budgets 20,25**	(Yes) Phase1-Window method performs same as Phase 2-model2, Variance method.(Budget=10)
For budgets 5,10,15, this phase 1 outperforms other two phase2 models. Overall, variance inferencing is best for these budgets.	Compared to other models in window inf, error % is least in this model for budgets 20,25.	
Varaince method of this model outperforms all 3 given models.		Hour based phase2 model out performs the other 3 models.
With each budget size increase of 5, error reduces to almost 50% for all 3 models.	With each budget size increase of 5, error reduces by almost 50% for all 4 models.	With each budget size increase of 5, error reduces to almost 50% for all 3 models.

Phase 3
Performance of window inferencing is better than variance based overall. For Budget 5,10, variance method performs better than window.
Compared to other models in variance inf, error % is least in this model for budgets 20,25.
With the increase in budget size abs error decreases.
Best results seen from this model, window based method, when taken highest budget(25)
For budgets 0,5, 10, phase 3 forward model performs better than the phase1 & 2 models.
For budgets 20,25 phase 2-hr model performs better than the other 3 models.

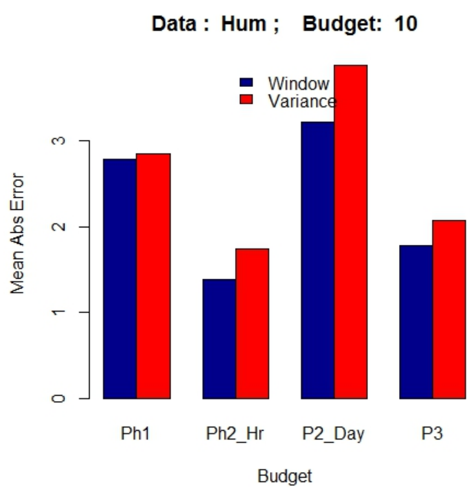
4.3 Humidity dataset plots for different budgets



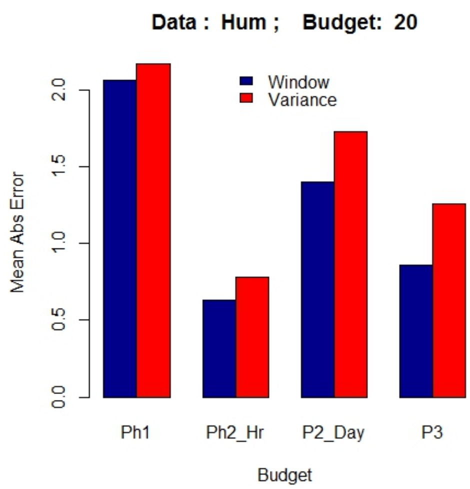
Plot 6- Phase1 , Phase2 , Phase3 models vs Mean Abs Error [Budget= 0]
(Phase1, Phase2 Model2, , Phase2 Model1, Phase3; Active inference (Window & Variance))



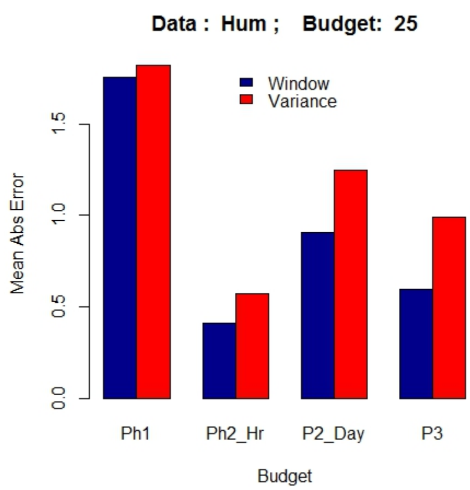
Plot 7- Phase1 , Phase2 , Phase3 models vs Mean Abs Error [Budget= 5]
(Phase1, Phase2 Model2, Phase2 Model1, Phase3; Active inference (Window & Variance))



Plot 8- Phase1 , Phase2 , Phase3 models vs Mean Abs Error [Budget= 10]
(Phase1, Phase2 Model2, , Phase2 Model1, Phase3; Active inference (Window & Variance))



Plot 9- Phase1 , Phase2 , Phase3 models vs Mean Abs Error [Budget= 20]
(Phase1, Phase2 Model2, Phase2 Model1, Phase3; Active inference (Window & Variance))



Plot 10- Phase1 , Phase2 , Phase3 models vs Mean Abs Error [Budget= 25]
(Phase1, Phase2 Model2 , Phase2 Model1, Phase3; Active inference (Window & Variance))

4.3.1 Comparison /Analysis (Humidity)

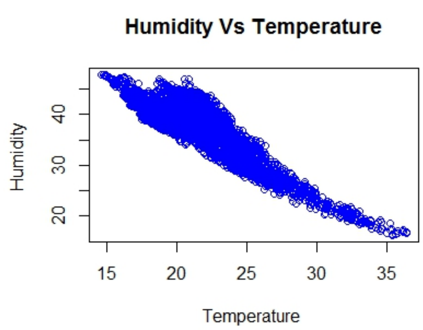
Looking at plots 1,2,3 for humidity datasets we conclude the following:

Phase 1	Phase2-Model 1 (hour)	Phase2-Model 2 (day)
Performance of Window inferencing is better than variance based.	**Best performance out of all 4 models.**	Performance of window inferencing is better than variance based.
With the increase in budget size absolute error decreases.	Performance of window inferencing is better than variance based.	With the increase in budget size abs error decreases.
Best results seen from: this model, Window based method, when taken highest budget(25)	Best results seen from: this model, Window based method, when taken highest budget(25)	Best results seen from: this model, Window based method, when taken highest budget(25)
(Yes) Phase1-Window method performs same as Phase 2-model1, Variance method.(Budget=5)	(Yes) Phase2-Window method performs same as Phase 2-model1, Variance method.(Budget=5)	Not similar
Phase1-Window method , Variance method, give almost same performance.		
For budgets 5,10,15 in this phase variance & window inferencing perform almost similar.	Compared to other two models individually, this model performs best.	For budgets 5,10,15, this phase model performs worst.
For budgets 20,25 this phase model performs worst.	Compared to other two models individually, this model performs best.	For budgets 20,25, this phase model performs better than the phase1 model.
	For all budgets, Window method of this model outperforms all 3 given models.	Hour based phase2 model out performs the other two models.
With each budget size increase of 5, error reduces to almost 50% for all 3 models.	With each budget size increase of 5, error reduces to almost 50% for all 3 models.	With each budget size increase of 5, error reduces to almost 50% for all 3 models.

Phase 3
Performance of Window inferencing is better than variance based.
With the increase in budget size absolute error decreases.
This is the second best model for humidity data.
Best results seen from: this model, Window based method, with highest budget(25)
Phase 3 results are comparable with Phase 2-model1
For budgets 5,10,15 in this phase variance & window inferencing perform almost similar.
For budgets 20,25 this phase model performs best.
With each budget size increase of 5, error reduces to almost 50% for all 4 models.

4.4 Temperature vs Humidity

As we can see from the plot below with the increase in temperature humidity decreases.



Plot 11- Temperature vs Humidity

5 Observations about sensors

Where betas are non-zero I considered it as value 1, and created an adjacency matrix of size 50×50 between the 50 sensors. Where there will be a correlation between sensors there will be a 1 in coresponding cell of sensor_i & sensor_j. I plotted this matrix in below given graph.

Figure 1: Nodes in dark blue color represent sensors, and edges are correlation between two sensors (if exists)

6 Technology used for the project

R version 3.2.4 , RStudio; MS Excel

OS- Windows 10

Packages- gdata, glmnet