Weather Data Hackathon Queries

Tom Dietterich, Eric Woldridge, Version 4, July 17, 2017

Introduction

This document provides the details on the hackathon queries and associated data files. Recall that the data is drawn from weather stations that lie along a transect of Oklahoma, as shown in Figure 1. Data from the yellow stations, ARNE, BUTL, FTCB, NINN, PAUL, and FITT, are provided only during the training period (4/1/2008-6/30/2008). Data from the red stations, LANE, ANTL, CLOU, and IDAB, are provided during both training and testing. The test data are drawn from the four green stations, KENT, BOIS, GOOD, and SLAP.

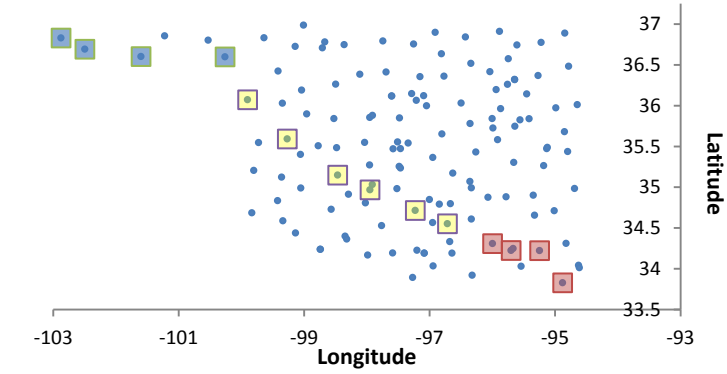

Figure 1. Locations of Oklahoma Mesonet weather stations. Stations along the transect are highlighted. Yellow: Training set only. Green: Test set only. Red: Available in both Train and Test sets.

Database Files

The training data (yellow and red stations) are provided in the sqlite3 database "train.db", which contains two tables:

The OBSERVATIONS table has the following columns:

Variable	Meaning	Units
STID	Station Identifier (4 letter code)	NA
DATE	Date and time of the observation in the format YYYY-MM-DD HH:MM:SS	UTC
RELH	Relative Humidity	%
TAIR	Air temperature @ 2m above surface	degrees C
WSPD	Average Wind Spd @ 10m above surface	m/s
WDIR	Wind Direction	deg from north
PRES	Barometric Pressure	mbar
SRAD	Solar Radiation	W/m^2
TSRAD	Theoretical Solar Radiation	W/m^2

Information about each weather station is provided in the GEOINFO table

Variable	Meaning	Units
stid	Station Identifier (4 letter code)	NA
nlat	North Latitude	degrees
elon	East Longitude	degrees
elev	Elevation above Sea Level	meters

The red station test data is provided in “test.db”. It has the same format as “train.db” .

We have also provided the same data in CSV format in the folders train-csvs and test-csvs in case that is more convenient.

Queries

Task 1: Coarse-Scale Prediction

The goal of this task is to predict summary statistics for each month. Specifically, you are to predict the monthly mean, high, and low temperature, mean, max, and min relative humidity, mean, max, and min solar radiation, mean, max, and min barometric pressure. You should make predictions for all four test stations. This query will be answered without any data from the test region stations. As output, please produce a comma-separated value file named “query1-predictions.csv” in the following format:

```
STID,MONTH,TAIR.MEAN,TAIR.MAX,TAIR.MIN,RELH.MEAN,RELH.MAX,RELH.MIN,SRAD.MEAN,SRAD.MAX,SR
AD.MIN,PRES.MEAN,PRES.MAX,PRES.MIN
KENT,07,--,--,--,--,--,--,--,--,--,--,--,--
KENT,08, -,--,--,--,--,--,--,--,--,--,--,--
KENT,09, -,--,--,--,--,--,--,--,--,--,--,--
BOIS,07, -,--,--,--,--,--,--,--,--,--,--,--
etc.
```

If you wish, you may also submit a set of samples from your posterior distribution. Each sample should be submitted as a file named “query1-predictions-nnn.csv” where “nnn” is replaced by a sample number starting from 000.

Task 2: Pure Prediction

Predict the temperature, pressure, relative humidity, and solar radiation for the KENT station. This query will be answered without any data from the test region stations. As output, please produce a comma-separated value file named “query2-predictions.csv” in the following format:

```
STID,DATE,TAIR,RELH,SRAD,PRES
KENT,"2008-07-01 00:00",--,--,--,--
KENT,"2008-07-01 00:05",--,--,--,--
KENT,"2008-07-01 00:10",--,--,--,--
…
KENT,"2008-09-30 23:55",--,--,--,--
```

If you wish, you may also submit a set of samples from your posterior distribution. Each sample should be submitted as a file named "query2-predictions-nnn.csv" where "nnn" is replaced by a sample number starting from 000.

Task 3: Conditional Prediction

For this task, we provide the sqlite3 database "query3.db". It contains the following OBSERVATIONS table:

Variable	Meaning	Units
STID	Station Identifier (4 letter code)	NA
DATE	Date and time of the observation in the format YYYY-MM-DD HH:MM:SS	UTC
RELH	Relative Humidity	%
TAIR	Air temperature @ 2m above surface	degrees C
WSPD	Average Wind Spd @ 10m above surface	m/s
WDIR	Wind Direction	deg from north
PRES	Barometric Pressure	mbar
SRAD	Solar Radiation	W/m^2

The observations are for the BOIS station. The observed values for PRES, SRAD, WSPD, and WDIR are provided, but the values for TAIR and RELH are given as "NA". Your task is to predict those missing values. (Note that no theoretical solar radiation, TSRAD, is provided, because it depends on the TAIR and RELH values.)

As output, please submit a sqlite3 database named "query3-predictions.db". This should be a copy of "query3.db" with the TAIR and RELH "NA" values replaced by your predictions.

If you wish, you may also submit a set of samples from your posterior distribution. Each sample should be submitted as a database file named "query3-predictions-nnn.db" where "nnn" is replaced by a sample number starting from 000.

For your convenience, we are also providing a database named "query3-train.db" which contains data for the five "yellow" stations with TAIR and RELH replaced by "NA". You can use this to test your prediction models and scripts.

Task 4: Imputation

This task is essentially identical to Task 3, except that the set of provided and missing variables is randomized. The observations are for the GOOD station. For this task, we provide the sqlite3 database "query4.db". It contains the following OBSERVATIONS table:

Variable	Meaning	Units
STID	Station Identifier (4 letter code)	NA
DATE	Date and time of the observation in the format YYYY-MM-DD HH:MM:SS	UTC
RELH	Relative Humidity	%
TAIR	Air temperature @ 2m above surface	degrees C
WSPD	Average Wind Spd @ 10m above surface	m/s
WDIR	Wind Direction	deg from north
PRES	Barometric Pressure	mbar
SRAD	Solar Radiation	W/m^2

In this database, approximate 40% of the readings for RELH, TAIR, PRES, and SRAD have been replaced by "NA". Your task is to predict these missing values.

As output, please submit a sqlite3 database named "query4-predictions.db". This should be a copy of "query4.db" with "NA" values replaced by your predictions.

If you wish, you may also submit a set of samples from your posterior distribution. Each sample should be submitted as a database file named "query4-predictions-nnn.db" where "nnn" is replaced by a sample number starting from 000.

For your convenience, we are also providing a database named "query4-train.db" which contains data for the five "yellow" stations with "NA" values inserted using the same proportions as in the query4.db. You can use this to test your prediction models and scripts.

Task 5: Quality Control

For this task, we provide an sqlite3 database "query5.db" with the following OBSERVATIONS table:

Variable	Meaning	Units
STID	Station Identifier (4 letter code)	NA
DATE	Date and time of the observation in the format YYYY-MM-DD HH:MM:SS	UTC
RELH	Relative Humidity	%
TAIR	Air temperature @ 2m above surface	degrees C
WSPD	Average Wind Spd @ 10m above surface	m/s
WDIR	Wind Direction	deg from

		north
PRES	Barometric Pressure	mbar
SRAD	Solar Radiation	W/m^2

The data are from the SLAP station. In this case, there are no missing values unless a sensor was not operating at a given time point. However, synthetic sensor failures have been injected into the TAIR, PRES, RELH, and SRAD fields at a rate of approximately 10%. Your task is to identify those sensor failures.

As output, please submit an sqlite3 database named "query5-predictions.db". This should contain the following ERRORS table in which each value indicates the probability that the sensor reading is bad. Hence, all good readings should have probabilities near zero and all bad readings should have probabilities near 1.

Variable	Meaning	Units
STID	Station Identifier (4 letter code)	NA
DATE	Date and time of the observation in the format YYYY-MM-DD HH:MM:SS	UTC
RELH	Relative Humidity	probability
TAIR	Air temperature @ 2m above surface	probability
PRES	Barometric Pressure	probability
SRAD	Solar Radiation	probability

Evaluation

Tasks 1-4 will be evaluated based on the squared error between the predicted and actual values. If the team submits posterior samples, then we will compute the posterior distribution of the squared error.

Task 5 will be evaluated based on area under the ROC curve. By sweeping a threshold through these scores, the evaluator will measure the AUC.

Summary of Files Provided

Name	Description
train.db	Training period data for yellow and red stations
test.db	Test period data for red stations
query3-train.db	Training period data for yellow stations with TAIR and RELH replaced by "NA" (for debugging)
query4-train.db	Training period data for yellow stations with various sensor readings replaced by "NA"

query3.db	Test period data for BOIS with TAIR and RELH replaced by “NA”
query4.db	Test period data for GOOD with various sensor readings replaced by “NA”
query5.db	Test period data for SLAP with synthetic sensor failures injected into TAIR, RELH, and PRES
Weather Data hackathon.pdf	This file

The query3.db, query4.db, and query5.db files will not be distributed until the teams agree that they are ready to receive them and make their predictions. This will probably take place around Noon on Tuesday.

Submitting Solutions

Solutions to each task are to be submitted in your teams private Gitlab repo provided by the TA1 team. Results for each task should be staged in the respective folder under the …\Results\ directory.

At 9:00 am EDT the TA1 team will pull all task solutions for evaluation.

Schedule

July 2017 PPAML Hackathon Schedule

Darpa Conference Center

Time	Mon 7/17	Tue 7/18	Wed 7/19	Thu 7/20
8:30 AM	Check-in		All Deliverables and Final Check-ins due (9:00 AM)	
9:00 AM	Hackathon Introduction	Team Out Briefs and Planning		
9:30 AM		Working Session	TA1 Evaluation	
10:00 AM	Model Brainstorm			TA1 Presentation Distributed
10:30 AM				
11:00 AM	Implementation Aproach by PPS			
11:30 AM				
12:00 PM	Lunch	Lunch		
12:30 PM				
1:00 PM				
1:30 PM	Working Session	Working Session		
2:00 PM				
2:30 PM				
3:00 PM				
3:30 PM				
4:00 PM				Hackathon Result Presentations (TA1 and TA2-4 Teams)
4:30 PM				
5:00 PM				
5:30 PM	Team Out Briefs	Out Briefs and Wrap-UP		