Machine Learning Assignment

Data Science Methods for Smart City Applications

Due: March 28, 2018

For this assignment, you will work with data from a large building "The Great Energy Predictor Shootout - The First Building Data Analysis and Prediction Competition" held in 1993-94 by ASHRAE. (For details see Behl, M., & Mangharam, R. (2014). Evaluation of DR-Advisor on the ASHRAE Great Energy Predictor Shootout Challenge.

https://repository.upenn.edu/cgi/viewcontent.cgi?article=1093&context=mlab_papers).

The training data was a time record of hourly chilled water, hot water and whole building electricity usage for a four-month period in an institutional building. Weather data and a time stamp were also included. The hourly values of usage of these three energy forms was to be predicted for the two following months. The testing set consisted of the two months following the four-month training period. The data had approximately 3000 samples taken hourly during Sep - Dec 1989. The following information was provided for each time step:

1. Outside temperature (F)
2. Wind speed (mph)
3. Humidity ratio (water/dry air)
4. Solar flux (W/m2)
5. Hour of Day
6. Whole building electricity, WBE (kWh/hr)
7. Whole building chilled water, CHW (millions of Btu/hr)
8. Whole building hot water, HW (millions of Btu/hr)

The corresponding dates are also provided.

Note that the input variables are: (1) Outside temperature (F); (2) Wind speed (mph); (3) Humidity ratio (water/dry air); and (4) Solar flux (W/m2). Note that these are all environmental variables.

For your assignment, consider the outcome variables to be: (1) whole building energy WBE, (2) chilled water consumption, CHW, and (3) hot water consumption HW.

You will construct predictive models for the outcome variables, using a subset (or all) of the input variables. You will run the following algorithms:

(1) Linear Regression
(2) Cubic Splines; and
(3) Support Vector Regression

In part 2 of the assignment, you will first pre-process the data, for example, you may run a clustering algorithm on the input data variables to see if they break down into different groups, corresponding to different environmental conditions. If you find the groupings formed by clustering to be meaningful, and then apply your predictive models for each group that you have obtained.

Use reasonable metrics, such as mean squared error (absolute fit) and r-squared (relative measure of fit), and methods, such as n-fold cross validation to compare the results you obtain from the different algorithms.

We recommend that you use the scipy library (https://www.scipy.org/) for your assignments. Documentation and code for the various algorithms can be found at:

For linear regression and splines:

https://docs.scipy.org/doc/scipy/reference/tutorial/interpolate.html. Also look up

https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9 (especially the section on Linear Regression in SKLearn).

https://docs.scipy.org/doc/scipy-0.18.1/reference/generated/scipy.interpolate.CubicSpline.html (cubic splines)

For support Vector Regression

http://scikit-learn.org/stable/modules/svm.html#svm-regression or

http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

For clustering, you may look up https://docs.scipy.org/doc/scipy/reference/cluster.html. In particular, for k-means clustering loo up:

http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

and for hierarchical clustering refer to:

https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html#module-scipy.cluster.hierarchy

(Suggestion for hierarchical clustering it is best to focus on the average algorithm, though you may also compare those results against complete link clustering.

Get an early start, and ask lots of questions. We know most of you are not familiar with these machine learning methods, but this is a good way to learn about them by applying them to data, and trying to interpret the results. Getting familiar with them should also help you with your group projects. The faculty and the graduate students (Avisek, Chinmaya) are here to help you as you work through this assignment. I will come back after Spring Break, and help you work through an example or two in class. Nevertheless, get started and see how much progress you can make on your own.