DATA 602 Fall 2016 Project Proposal, Part 2

James Topor

November 10, 2016

Introduction

Sensor technology is rapidly changing the ways in which electricity consumption can be monitored and controlled while also improving the effectiveness of automated security systems. For example, home thermostats and lighting can now be programmed to activate according to the amount of daylight detected via an external light sensor, thereby ensuring that lighting and/or HVAC systems are deactivated when not needed. Motion sensors have also long been used in many commercial settings to control the activation and deactivation of lighting systems. However, a common shortcoming of motion detection systems is that they will deactivate a lighting system if they fail to detect sufficient motion within the space they are monitoring. As a result, individuals can find themselves suddenly cast into darkness while working at a desk or while using various public facilities.

Additional types of sensor technology are rapidly evolving to allow for their application to both energy management and security within both commercial and residential structures. As these technologies become both less expensive to use and easier to implement, building owners and energy management firms are exploring ways in which multiple types of sensors can be combined to more accurately control energy usage and/or improve building security.

Research Question

As the use of additional types of sensors becomes practical on a widespread basis, a potential research question that can be posed is:

- Can the output from multiple types of sensors allow us to determine whether or not people are present within a room?

If so, both energy management and security systems might benefit from the use of such sensors. For example, energy management systems could be refined so as to minimize the likelihood of a lighting system being turned off due to minimal motion of a room's occupants. Similarly, security systems could be further enhanced to allow for better detection of potential intruders.

Two researchers explored this question in a paper published earlier this year: ("**Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models**". Luis M. Candanedo, Varonique Feldheim. Energy and Buildings. Volume 112, 15 January 2016, Pages 28-39).

The goal of this DATA 602 project is to apply some of Python's regression tools in an attempt to predict the occupancy of a room based on the output of a variety of different types of sensors.

Data To Be Used

The data for this project will be sourced from the University of California at Irvine's Machine Learning Repository. Specifically, we will make use of the **Occupancy Detection Data Set** accessible via the following web link:

- https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

The data set is comprised of a total of 20,560 cases spread across three files, with each case being comprised of six attributes and one response variable:

Date: A 19-character field comprised of the date and time of the recording of the case. The 19 character field is structured as follows:

```
YYYY-MM-DD HH:MM:SS
```

where YYYY is the year (4 digits), MM is the month (2 digits), DD is the day (2 digits), HH is the hour (2 digits), MM is the minute (2 digits), and SS is the second (2 digits)

Temperature: The recorded room temperature in degrees Celsius.

Relative Humidity: The percentage relative humidity recorded within the room.

Light: The amount of light recorded within the room, expressed in terms of "lux". A lux equals one lumen incident per square meter of illuminated surface area.

CO2: The amount of carbon dioxide recorded within the room, measured in parts per million (ppm).

Humidity Ratio: A derived metric measured as (kilograms of water vapor / kilograms of air).

Occupancy: A binary 0/1 flag, with 0 indicative of the room not having been occupied and 1 indicative of the room having been occupied at the time of the measurements. This field represents the response variable for the observation.

The three separate data files can be viewed in their entirety at the following Github links:

- https://raw.githubusercontent.com/jtopor/CUNY-MSDA-602/master/Project/datatest.txt
- https://raw.githubusercontent.com/jtopor/CUNY-MSDA-602/master/Project/datatest2.txt
- https://raw.githubusercontent.com/jtopor/CUNY-MSDA-602/master/Project/datatraining.txt

Two of the provided data files (datatest.txt and datatest2.txt) will be used for model building/training while the third (datatraining.txt) will be used for model testing.

Approach

The project will be created using iPython Notebook and will be comprised of 4 distinct components: Data Preparation, Data Exploration, Regression Modeling, and Model Selection. Each one is described below.

Data Preparation

The data obtained via the UCI website may need to be refined somewhat for purposes of this project. For example, data fields may require reformatting and/or transformation for purposes of enabling the intended analysis. Data Preparation will address any such issues and will result in the creation of two separate data frames (one containing data to be used for model training and one containing data to be used for model testing) that are fully prepped for data exploration, regression modeling, and model selection.

Data Exploration

Data Exploration will include detailed analysis of the various variables contained within the data set via summary statistics, histograms, boxplots, and correlation matrices. Results of Data Exploration will be used to determine whether any variables should be dropped from the analysis due to collinearity and will also serve as the basis for determining whether any of the predictor variables may need to be transformed as part of the model building process.

Regression Modeling

Functions contained within Python's **statsmodels** package will be used to construct two different binary logistic regression models as well as one linear model of the relationship between the **occupancy** response variable and the five independent predictor variables (Temperature, Relative Humidity, Light, CO2, and Humidity Ratio). Model building will rely exclusively on a training set of data obtained from the UCI website. Each model will be refined so as to minimize collinearity amongst the predictor variables and relevant model fit and performance metrics will be gathered and discussed.

The **logit** function contained within the **statesmodels** package will be used for constructing the binary logistic regression models while the **ols** function from that package will be used for generating the linear regression model. Each model will be assessed via model fit metrics calculated by the **logit** and **ols** functions as well as via performance metrics as calculated by functions contained within the **sklearn.metrics** package. Specifically, the **accuracy_score**, **average_precision_score**, **roc_auc_score**, **f1_score**, and **recall_score** functions will be applied to the results of each model to assess model performance. Also, when possible, model fit graphics such as Q-Q plots and added variable plots will be generated to assist in the assessment of model fit.

Finally, a discussion of the inferences we can make based on the directionality of the resulting regression model coefficients for each independent variable will be provided.

Model Selection

The three models will then be compared against one another via model fit and performance metrics such as log likelihood, R^2, accuracy, precision, sensitivity, and F-statistics. The "best" model will then be applied to the testing data set obtained from the UCI web site. The results of that testing will then be compared against the actual occupancy indicator flag via the aforementioned performance metrics to allow us to evaluate the efficacy of the selected model.

Conclusion

Lastly, a brief discussion of any conclusions we can draw from the work described herein will be provided, including suggestions for possible further research.