OCCUPANCY DETECTION

A Project Report Submitted by:

Jinto Jose (ID: 01677777)

Abstract

Corporate companies or firms in general want to make their offices likeable to their employees so that they can get more productivity out of them. They also want the offices to be likeable to their clients so that they can get more business out of them. One of the ways to do this is to find out the best time suitable as well as the best circumstances for the person or people to be comfortable in a room. This can be done by finding out various factors in the room when the room is occupied or not occupied. After these details are collected, the occupancy can be analyzed using various Machine Learning algorithms. The most accurate Machine learning algorithm will predict the most appropriate occupancy results.

Introduction

The goal of this project would be to correctly predict the occupancy of a room for certain parameters pertaining to that room. This is important since based on these parameters, the company/firm can decide what parameters to use for an ideal work environment to be present in a room.

The dataset being used for this study is the “Occupancy Detection Data Set” from the reputed UCI Machine Learning Repository. Following are the parameters taken into consideration in the data set: -

1. Temperature
2. Humidity
3. Light
4. CO_2
5. Humidity Ratio

There are a total of 3 datasets provided by the repository – One for training and the other two are for testing. The training dataset consists of the above 5 parameters as well as a 6^{th} field which is Occupancy which is denoted as 0 or 1 based on whether the room is not occupied or occupied. The following algorithms were trained with the training dataset and this data was used to test the test dataset: -

1. Logistic Regression
2. Linear Discriminant Analysis

3. Naïve Bayes Classifier
4. Quadratic Discriminant Analysis

The accuracy of each of these algorithms was found out on the test datasets and compared to each other to find out which was the best approach to find out the occupancy.

Background

Not much work had been put into this topic as this is a fresh topic. Current approaches to occupancy detection take place mostly in commercial buildings using passive infrared (PIR) motion detectors. However, motion detectors have inherent limitations when occupants remain relatively still. The use of probabilistic models offers improved capability of detecting occupant presence. However, the fundamental dependence on motion still remains. Moreover, motion detectors alone only provide information regarding the presence or absence of people in a space rather than the number of occupants.

Video cameras have been used in this regard; however, video capture raises privacy concerns and requires large amounts of data storage. Other work has focused on the use of carbon dioxide (CO2) sensors in conjunction with building models for estimating the number of people generating the measured CO2 level. Sufficient models, though, are often not easy to obtain and extensions to complex or open spaces may be difficult. In general, occupancy detection that fully exploits information available from low cost, non-intrusive, environmental sensors is an important yet little explored problem in office buildings.

Approach

The following algorithms were used to test out the datasets:-

1. Logistic Regression

   In statistics, logistic regression is a regression model where the dependent variable (DV) is categorical. This is the case of a binary dependent variable—that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail, win/lose, alive/dead or healthy/sick.

   Logistic regression was developed by statistician David Cox in 1958. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the probability of a given outcome by a specific percentage.

We used multiple programs in Octave to implement this algorithm. We used a cost function, sigmoid, fminunc(for gradient descent) and a main program. Sigmoid function is as follows: -

$$g(z) = \frac{1}{1+e^{-z}}$$

The cost function is as follows: -

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))]$$

2. Linear Discriminant Analysis

LDA is a generative model for classification that assumes the class covariances are equal. Given a training dataset of positive and negative features (x, y) with $y \epsilon \{0, 1\}$. LDA models the data x as generated from class-conditional Gaussians:

$$P(x, y) = P(x|y)P(y) \; where \; P(y = 1) = \pi \; and \; P(x|y) = N(x; \mu^y, \Sigma)$$

where means μ^y are class-dependent but the covariance matrix is class-independent (the same for all classes). A novel feature x is classified as a positive if $P(y = 1|x) < P(y = 0|x)$, which is equivalent to $a(x) > 0$,where the linear classifier $a(x) = w^Tx + w_0$ has weights given by

$$w = \underline{\Sigma}^{-1}(\mu^1 - \mu^0)$$

In practice, we use a(x) > some threshold, or equivalently, $w^Tx > T$ for some constant T.

LDA works when the measurements made on independent variables for each observation are continuous quantities. When dealing with categorical independent variables, the equivalent technique is discriminant correspondence analysis. LDA explicitly attempts to model the difference between the classes of data.

3. Naive Bayes Classifier

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features. With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

Naive Bayes classifiers are highly scalable, requiring several parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

$$\hat{\mathbb{P}}(y = j|x_0) = \frac{\hat{\pi}_j \hat{f}_j(x_0)}{\sum_{k=1}^{K} \hat{\pi}_k \hat{f}_k(x_0)}$$

$$\text{logit}(y = 1|\mathbf{x}) = \beta_0 + \sum_{k=1}^{K} g_k(x_k)$$

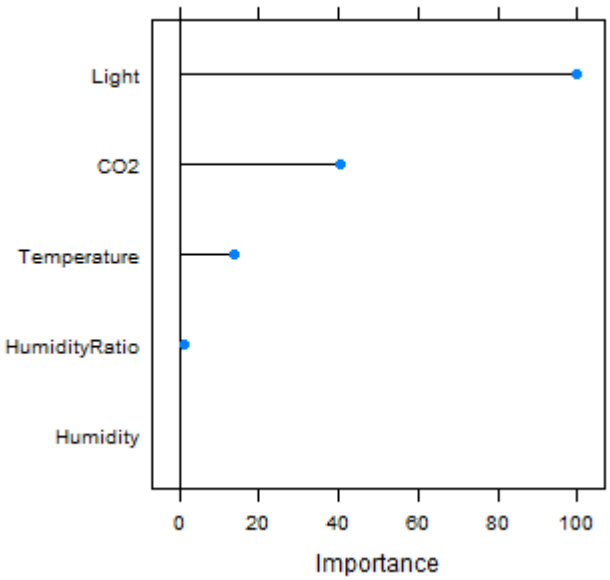
4. <u>Quadratic Discriminant Analysis</u>

Quadratic discriminant analysis (QDA) is closely related to linear discriminant analysis (LDA), where it is assumed that the measurements from each class are normally distributed. Unlike LDA however, in QDA there is no assumption that the covariance of each of the classes is identical. When the normality assumption is true, the best possible test for the hypothesis that a given measurement is from a given class is the likelihood ratio test. Suppose there are only two groups, (so $y \, \epsilon \, \{0, 1\}$), and the means of each class are defined to be $\mu_{y=0}$, $\mu_{y=1}$ and the covariances are defined as $\Sigma_{y=0}, \Sigma_{y=1}$. Then the likelihood ratio will be given by

$$\text{Likelihood ratio} = \frac{\sqrt{|2\pi\Sigma_{y=1}|}^{-1} \exp\left(-\frac{1}{2}(x-\mu_{y=1})^T \Sigma_{y=1}^{-1}(x-\mu_{y=1})\right)}{\sqrt{|2\pi\Sigma_{y=0}|}^{-1} \exp\left(-\frac{1}{2}(x-\mu_{y=0})^T \Sigma_{y=0}^{-1}(x-\mu_{y=0})\right)} < t$$

Results

From the dataset description, the order of importance of the parameters used was given as follows: -



Following were the results obtained from each of the algorithms on the two test datasets: -

1. Logistic Regression

Code was written in Octave and the accuracies are as follows: -

Test data set 1 Accuracy: 97.523452
Test data set 2 Accuracy: 98.205496

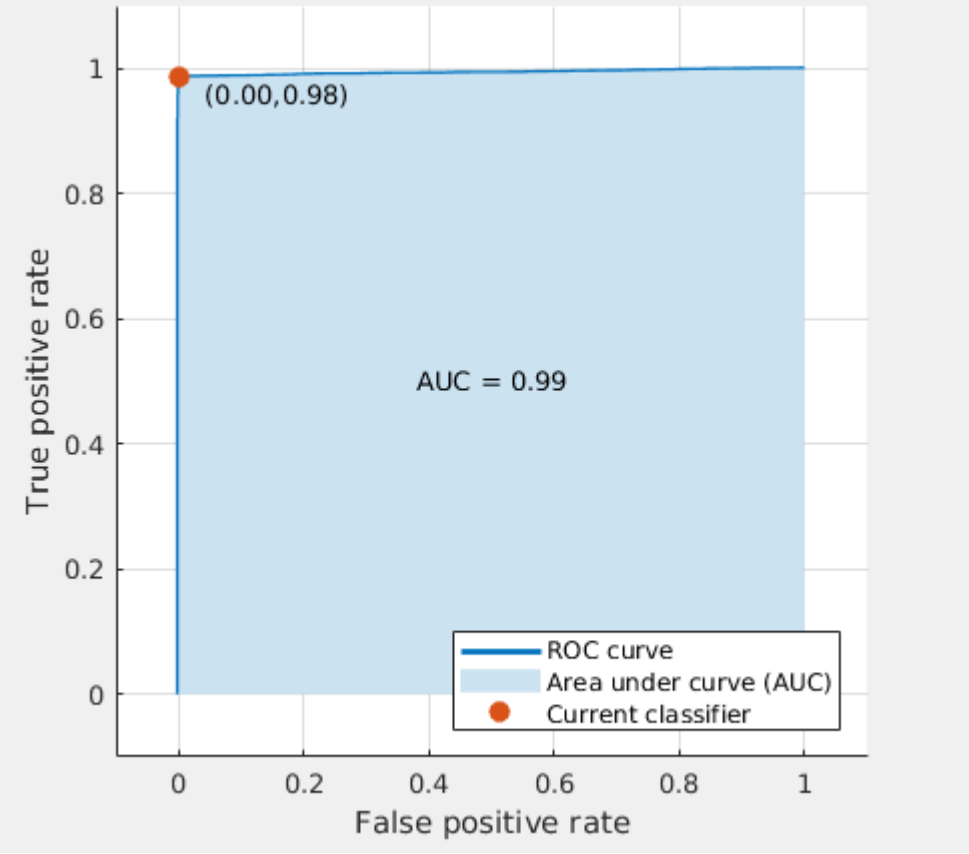
2. Linear Discriminant Analysis

Code was written in Octave and the accuracies are as follows: -

Test data set 1 Accuracy: 97.898687
Test data set 2 Accuracy: 98.759229

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings

The ROC curve for LDA is shown below: -



3. Naïve Bayes Classifier

Code was written in MATLAB and the accuracies are as follows: -

Test data set 1 Accuracy: 94.2964
Test data set 2 Accuracy: 95.9906

A confusion matrix, also known as an error matrix is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Confusion matrices for the datasets are as follows: -

confusion matrix for test set 1:

1648	45
107	865

confusion matrix for test set 2:

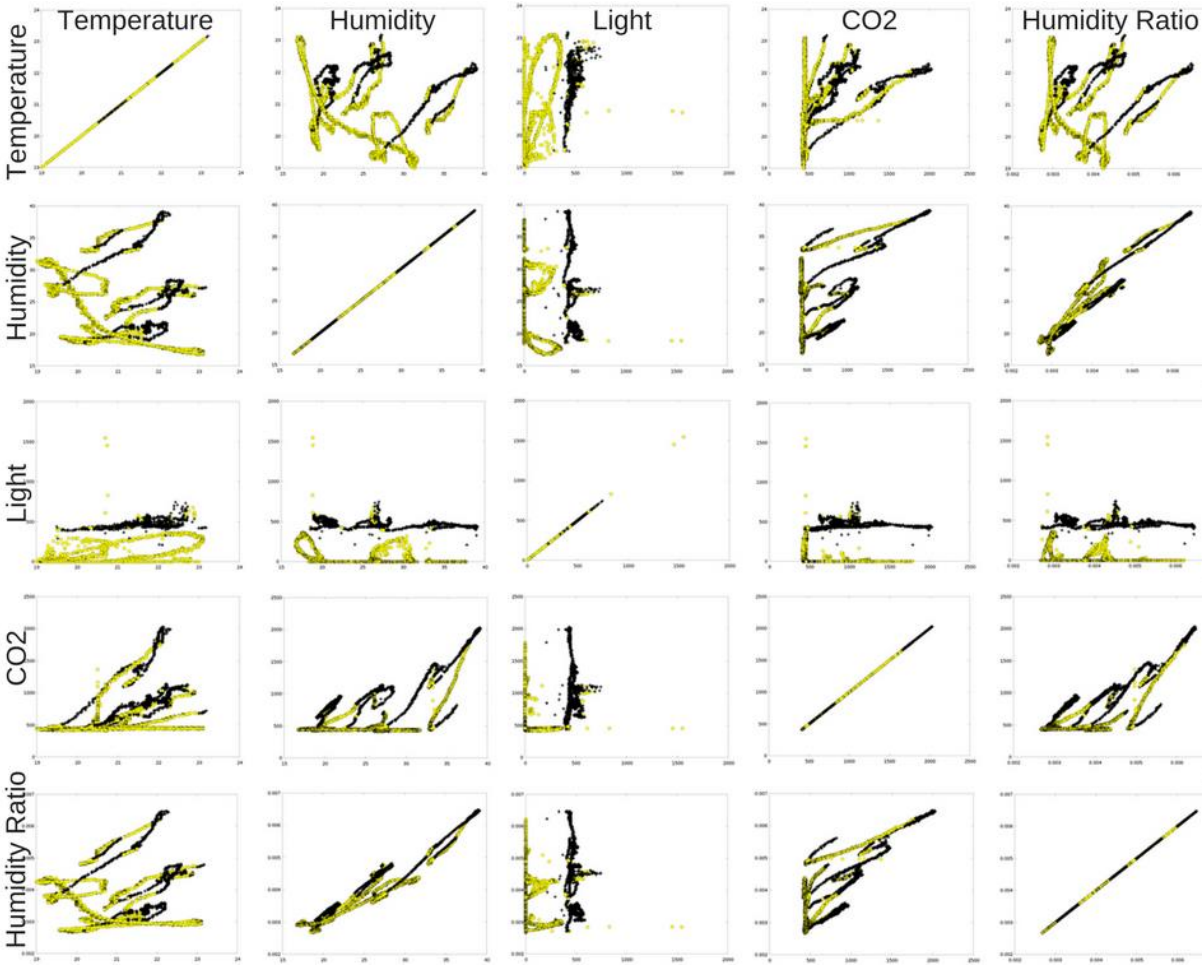
7643	60
331	1718

4. Quadratic Discriminant Analysis

Code was written in Octave and the accuracies are as follows: -

Test data set 1 Accuracy: 97.748593
Test data set 2 Accuracy: 98.677194

Plotting data for positive and negative classes with dataset features: -



This project can also be done using Support Vector Machines, Neural Networks and Hidden Markov Models.

Conclusion

From the results obtained, we could conclude that the best possible algorithm for the occupancy detection dataset is Linear Discriminant Analysis or LDA. Although Logistic Regression and Quadratic Discriminant Analysis came close to reasonable predictions. Comparison of Accuracies according to different algorithms: -

Algorithms	Test Dataset 1	Test Dataset 2
Logistic Regression	97.523452	98.205496
Linear Discriminant Analysis	97.898687	98.759229
Naïve Bayes Classifier	94.2964	95.9906
Quadratic Discriminant Analysis	97.748593	98.677194

References

1. https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+#
2. Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models. Luis M. Candanedo, VÃ©ronique Feldheim. Energy and Buildings. Volume 112, 15 January 2016, Pages 28-39.
3. https://en.wikipedia.org/wiki/Logistic_regression
4. https://en.wikipedia.org/wiki/Linear_discriminant_analysis
5. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
6. https://en.wikipedia.org/wiki/Quadratic_classifier
7. https://en.wikipedia.org/wiki/Confusion_matrix