Occupancy Detection

MT7038

Anton Stråhle & Max Sjödin

Fall 2020

Data

The occupancy status of a room was observed for a few days. Snapshots of the features below were taken every minute.

- Features
  - Temperature
  - CO2
  - Humidity
  - HumidityRatio
  - Light
- Response
  - Occupancy
    - Occupied
    - Unoccupied

Data

The occupancy status of a room was observed for a few days. Snapshots of the features below were taken every minute.

- Features
  - Temperature
  - CO2
  - Humidity
  - HumidityRatio
  - ~~Light~~
- Response
  - Occupancy
    - Occupied
    - Unoccupied

Light is excluded as the best classifier would otherwise become *Are the lights on?*

Brief Exploration

- Unbalanced data set
  - Many more unoccupied data points than occupied

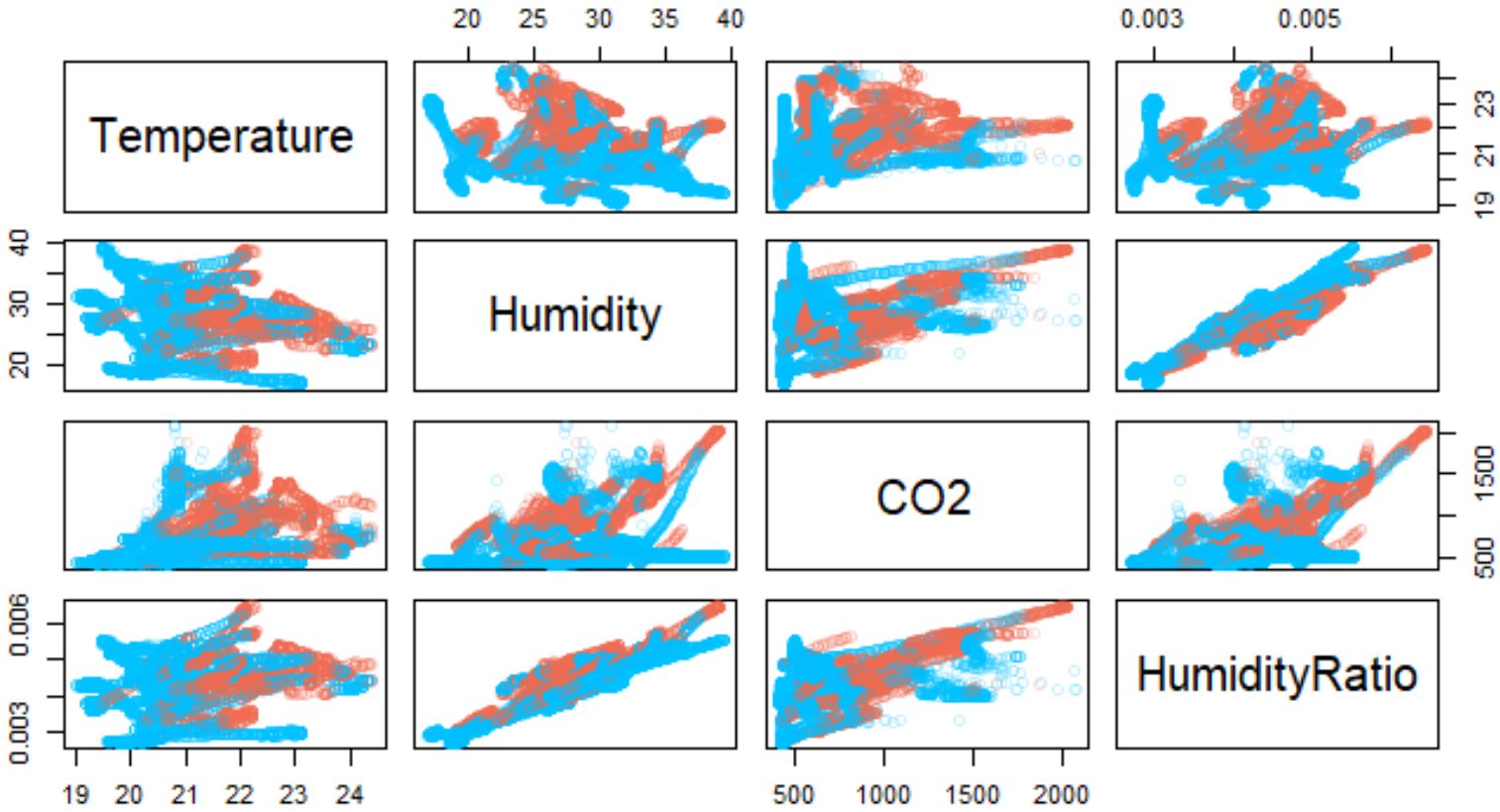

Figure: Pairplots of Features

- Non-linearity?

Brief Exploration

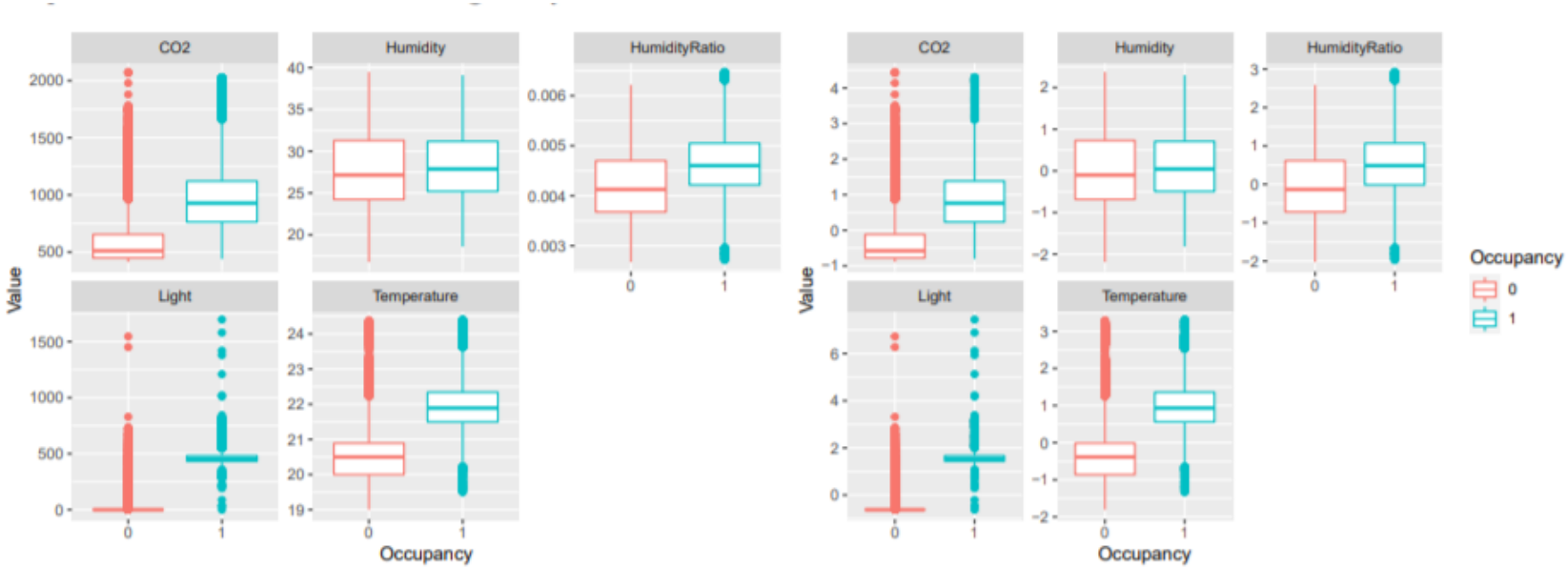

Figure: Boxplots of Features: Standardized and unstandardize

Methodology

- SVM
  - Linear, Radial & Polynomial
- KNN
  - Regular & Weighted
- Decision Trees
  - Single Tree, Bagging, Boosting

Methodology

SVM

Why? ▷ Good for non-linear classification problems.

How? ▷ Using the package **e1071** and the function **svm**
▷ Linear, polynomial and radial kernels
▷ Coarse-to-fine parameter search

Kernel	Cost	TestAccuracy
Linear	10.000	0.837
Radial	31622.777	0.915
Polynomial degree 5	316.228	0.922

Figure: SVM Accuracies

Thoughts? ▷ High costs $\Longrightarrow$ favoring low bias and perhaps an underlying lack of noise.

Methodology

KNN

Why?
- Good for non-linear classification problems
- Good with large training data sets
- Good if data is not noisy

How?
- Regular KNN using the package **class** and the function **knn**
- Weighted KNN using the package **kknn** and the function **kknn**
  - Epanechnikov kernel

Methodology

KNN - Regular

After a coarse search for a good value of k we noted that the best classifier was a 1-NN which further indicates that the data is not very noisy at all. The 1-NN achieved a testing accuracy of 93%.

A possible improvement is to use a weighted KNN where we put more emphasis on training points closer to the point which we want to predict than those further away.

Methodology

KNN - Weighted

In order to weight our data points we use the kernel distance from the point we want to predict to the k nearest neighbours. The choice of kernel is of course important but in our case all the available kernels in the function **kknn** generated approximatley the same results. As such we resorted to the Epanechnikov kernel as it is one we've encountered before.

When search for a good value of k in this case we found that the best validation accuracy was obtained for $k = 5$ which seems a bit more stable than using a 1-NN. This was also reflected in the testing accuracy which turned out to be 98.4%.

Methodology

Decision Trees

Why?
- ▷ Good with non-linear data
- ▷ Good with binary classification

How?
- ▷ Simple decision trees using package **rpart** using function **rpart**
- ▷ Bagging using the package **ipred** using function **bagging**
- ▷ Boosting using the package **adabag** using function **boosting**

Methodology

Decision Trees - Single

Training a single decision tree resulted in the high test accuracy of *xxx*. This in itself is a strong classifier but it can be further optimized using the aforementioned methods called bagging and boosting.

Bagging algorithms create B bootstrap samples, each samples is fitted by a decision tree. These trees are then averaged to create the so-called bagged tree.

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b^*(x)$$

The bagged tree however is no longer a tree. This unfortunately means that the interpretability of the decision tree is lost.

Methodology

Decision Trees - Boosting

Boosting algorithms iterate over a number of trees and for each iteration weight all observations. Correctly classified observations recieve less weight and incorrectly classified observations recieve more weight. This results in each iteration concentrating more and more on the incorreclty classified observations

Discussion

TABLE OF ALL TREES AND IMORTANT THINGS FOUND EARLIER(MODEL TEST ACCURACIES)
We concluded that bagging of decision trees resulted in the highest test accuracy out of the models tested. SOMETHING ABOUT THE LOW BIAS OF TREES AND THE VARIANCE REDUCTION OF BAGGING.