# Room Occupancy Detection Using Environmental Sensors

*A thesis is submitted to the department of*

***Computer Science & Engineering***

*of*

***International Institute of Information Technology Bhubaneswar***

*in partial fulfilment of the requirements*

*for the degree of*

***Bachelor of Technology***

*by*

***Abhishek Sharma***

*(Roll- B116002)*

*&*

***Sai Tharun***

*(Roll- B116041)*

*under the supervision of*

***Prof. Sabyasachi Patra***


**Computer Science**
**International Institute of Information Technology Bhubaneswar**
**Bhubaneswar Odisha - 751003, India**
**2020**


**International Institute of Information Technology Bhubaneswar**

Bhubaneswar Odisha -751 003, India. www.iiit-bh.ac.in

May 25, 2020

# Undertaking

We declare that the work presented in this thesis titled An Analysis Of The Occupancy Detection Dataset, submitted to the Department of Computer Science and Engineering, International Institute of Information Technology, Bhubaneswar, for the award of the Bachelors of Technology degree in the Computer Science and Engineering, is our original work. We have not plagiarised or submitted the same work for the award of any other degree. In case this undertaking is found incorrect, we accept that our degree may be unconditionally withdrawn.

***Abhishek Sharma***

***B116002***

***Sai Tharun***

***B116041***


**International Institute of Information Technology Bhubaneswar**
Bhubaneswar Odisha -751 003, India. www.iiit-bh.ac.in

May 25, 2020

# Certificate

This is to certify that the work in the thesis entitled *Room Occupancy Detection Using Environmental Sensors* by *Abhishek Sharma* and *Sai Tharun* is a record of an original research work carried out by them under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of *Bachelor of Technology* in *Computer Science Engineering*. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

***Sabyasachi Patra***
***Assistant Professor, Computer Science***
***IIIT, Bhubaneswar***

# Acknowledgment

The elation and gratification of this seminar will be incomplete without mentioning all the people who helped us to make it possible, whose gratitude and encouragement were invaluable to us. We would like to thank God, almighty, our supreme guide, for bestowing his blessings upon us in our entire endeavor. We express our sincere gratitude to Prof. Sabyasachi Patra , for his guidance and support and students of our class for their support and suggestions.

***Abhishek Sharma***

***B116002***

***Sai Tharun***

***B116041***

# Abstract

Detecting room occupancy accurately is an important research problem for reducing electricity consumption and combating security breaches. However, in cases where privacy concerns exist, detecting room occupancy without the use of surveillance cameras is preferred. The purpose of this project was to detect whether a room is occupied or not, solely from environmental sensor readings. To this end, we have used two statistical methods, namely Gaussian Mixture Models(GMM) and Gaussian Hidden Markov Models(GHMM), to detect room occupancy from sensor readings of Temperature, Humidity, Light and CO2 levels found in the Occupancy Detection Dataset. By naively applying said statistical models, we achieve an accuracy of 85%. We then compare the effectiveness of both statistical methods and question the idea of treating the Occupancy Detection Dataset as a time-series dataset. Finally, using a few useful observations, we will increase the accuracy of GMMs to 97%.


***Keywords*:** Occupancy Detection Dataset, Gaussian Mixture Models, Gaussian Hidden Markov Models.

# Contents

**Abstract** **vi**

**1 Gaussian Mixture Model** **1**
1.1 **The general framework of Mixture Models** 1
1.1.1 Generative point of view of mixture models 3
1.1.2 Probablistic point of view of mixture models 3
1.2 **The framework of GMMs** . . . . . 3
1.2.1 The setup . . . . . 3
1.2.2 Estimating the parameters of a GMM from data . . . . . 4
EM on GMM . . . . . 4

**2 Hidden Markov Model** **6**
2.1 Notation . . . . . 6
2.2 Assumptions . . . . . 7
2.3 Central Issues in HMMs . . . . . 7
2.4 Solution of the first central issue in HMMs . . . . 8
2.4.1 Forward algorithm . . . . . 8
2.4.2 Backward algorithm . . . . . 8
2.5 Solution of the second central issue in HMMs . . 8
Viterbi algorithm . . . . . 8
2.6 A few necessary definitions . . . . . 9
2.7 Solution of the Third central issue in HMMs . . . 9

**3 Analysis Of The Occupancy Detection Dataset :** **11**

**4 Predictions** **13**
4.1 A useful observation . . . . . 13

**Appendices** **15**

**A Appendix** **16**
A.1 Expectation-Maximization algorithm . . . . . 16

# Chapter 1

# Gaussian Mixture Model

The "generative model" –for data– need not follow a simple probability distribution. For example, the model generating our data might be multi-modal. In such cases, we might want to model this "generative model" as a *mixture* of simple probability distrbutions.Usually, this simple probability distribution is chosen as the gaussian distribution, due to its many convenient analytical properties.

## 1.1 The general framework of Mixture Models

The mathematical formulation of the generative model(of data) as a mixture model, is done in terms of *latent variables* and *observables*. Latent variables correspond to the mixture component( a.k.a cluster-labels) and observables correspond to the data-points. Latent variables will be denoted by $y$ and observables will be denoted by $x$. It is usually assumed that $y \in \{1, 2, 3, ..., M\}$ and $x \in \mathbb{R}^d$

In general, a mixture model assumes that the data-points are generated by the following process.

1. sample y (i.e. select the cluster)

2. sample x from a distribution that depends on y (i.e. sample

a point from selected cluster)

It then follows that $p(y, x) = P(y) * p(x|y)$. Here, $P(y)$ is *always* a multinomial distribution, while $p(x|y)$ can take a variety of parametric forms.
Let $y \sim Multinomial(\alpha)$, where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_M]$.
It follows then (from total probability theorem) that

$$p(x) = \sum_y P(y) * p(x|y) = \sum_{i=1}^{M} \alpha_i * p_i(x|y = i)$$

When, $x|y = i \sim \mathcal{N}(\mu_i, \Sigma_i)$, we call the mixture model a Gaussian Mixture Model(GMM).
*Prior* probability that a data point $x_i$ belongs to the k-th cluster is

$$P(y_i = k) = \alpha_k$$

*Posterior* probability that a data point $x_i$ belongs to the k-th cluster is

$$\begin{aligned} P(y_i = k|x_i) &= \frac{p(y_i = k, x_i)}{p(x_i)} \\ &= \frac{P(y_i = k)p_i(x_i|y_i = k)}{\sum_{j=1}^{M} P(y_i = j)p_i(x_i|y_i = j)} \\ &= \frac{\alpha_k p_i(x_i|y_i = k)}{\sum_{j=1}^{M} \alpha_j p_i(x_i|y_i = j)} \end{aligned}$$

$P(y_i = k|x_i)$ is an important quantity. It is also called the "responsibility" of component $k$ in generating data point $x_i$, and is more commonly denoted by $\gamma_{ik}$.
i.e

$$\gamma_{ik} = P(y_i = k|x_i)$$

By using the concept of responsibilities, we can view mixture models from two different points of view.

### 1.1.1 Generative point of view of mixture models

There are $M$ clusters, and each data point is generated by one of the M clusters. Hence, here $\gamma_{ik}$ is the probability that $x_i$ was *generated* by the k-th cluster

### 1.1.2 Probablistic point of view of mixture models

There are $M$ clusters, and each data point belongs to each clusters upto a certain extent. Hence, here $\gamma_{ik}$ is the extent to which $x_i$ belongs to cluster $k$. i.e. from a probabilistic point of view, mixture models perform 'soft' clustering on the dataset.

## 1.2 The framework of GMMs

### 1.2.1 The setup

data point is denoted by $x \in \mathbb{R}^d$.
cluster-label or mixture-component for x is denoted by $y$.
$y \sim Multinomial(\alpha)$, where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_M]$
i.e.

$$P(y = i) = \alpha_i$$

The pdf of the i-th gaussian distribution with parameters $\theta_i = \{\mu_i, \Sigma_i\}$ is

$$\begin{aligned} p_i(x_i|y_i = k) &= \underline{p_i(x_i|\theta_k)} \\ &= \mathcal{N}(x|\mu_k, \Sigma_k) \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} exp\{\frac{1}{2}(x_i - \mu_k)^T|\Sigma_k|^{-1}(x_i - \mu_k)\} \end{aligned}$$

Hence, by applying the total probability theorem

$$p(x_i) = \sum_{k=1}^{M} \alpha_k * p_i(x_i|y_i = k) = \sum_{k=1}^{M} \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

Let the set of all parameters of GMM be denoted by $\Theta$.
i.e.

$$\Theta = \{\alpha_1, \alpha_2, ..., \alpha_M, \theta_1, \theta_2, ..., \theta_M\}$$

### 1.2.2 Estimating the parameters of a GMM from data

Given a dataset(where each datapoint is in $\mathbb{R}^d$), to fit a GMM to it, we need to estimate

1. M = number of mixture-components
2. $\alpha_i, \forall i \in \{1, 2, ..., M\}$
3. $\mu_i \in \mathbb{R}^d, \forall i$
4. $\Sigma_i \in \mathbb{R}^{d*d}, \forall i$

We will assume that $M$ is already known.

#### EM on GMM

We will be using the Expectation-Maximization algorithm for estimating the parameters of the Gaussian Mixture Model.

Using the EM algorithm, the estimated parameters turn out to be:

$$\alpha_l^{new} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{il}$$

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} x_i \gamma_{il}}{\sum_{i=1}^{N} \gamma_{il}}$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} \gamma_{il}(x_i - \mu_l^{new})(x_i - \mu_l^{new})^T}{\sum_{i=1}^{N} \gamma_{il}}$$

where, $\gamma_{il} = p(l|x_i, \Theta^g)$

# Chapter 2

# Hidden Markov Model

Hidden Markov Model(HMM) is a doubly stochastic model, that is built on top of a markov chain. HMMs are used to model systems, where

1. states of the system can be modelled using a markov chain
2. states of the system can not be observed directly
3. emmisions from the states can be observed

## 2.1 Notation

$Q_t$ = hidden state at time $t$, $Q_t \in \{1, 2, ..., N\}$
$Q_t$ is a random variable.
$Q = (q_1, q_2, ..., q_T)$ = a *particular* sequence of states
$q_t \in \{1, 2, ..., N\}$ is a number

$O_t$ = observation at time $t$, $O_t \in V = \{v_1, v_2, ..., v_L\}$
$O_t$ is a random variable.
$O = (o_1, o_2, ..., o_T)$ = a *particular* observation sequence
$o_t \in V = \{v_1, v_2, ..., v_L\}$ is a number

$\pi = [\pi_1, \pi_2, ..., \pi_N]$, where $\pi_i = P(Q_1 = i)$

$a_{ij} = P(Q_t = j | Q_{t-1} = i)$
$A = [a_{ij}]_{N*N}$ is a row stochastic matrix.

$b_j(o_t) = p(O_t = o_t | Q_t = j)$
$B = [b_j(o_t)]_{N*T}$

$\lambda = [A, B, \pi]$ = the set of all parameters of the HMM

## 2.2 Assumptions

There are two assumptions that HMMs obey. They are:

1. $\{Q_{t:T}, O_{t:T}\} \perp\!\!\!\perp \{Q_{1:t-2}, O_{1:t-1}\} | Q_{t-1}$
2. $X_t \perp\!\!\!\perp \{Q_{\neg}, X_{\neg t}\} | Q_t$

## 2.3 Central Issues in HMMs

1. Given the parameters of the model, find the probability of observing a particular emmision sequence.
   i.e. find
   $$p(O|\lambda)$$
2. Given the parameters of the model, find the most likely sequence of states.
   i.e.find
   $$\underset{Q}{\operatorname{argmax}} P(Q|O, \lambda)$$
3. find
   $$\lambda^* = \underset{\lambda}{\operatorname{argmax}}\, p(O|\lambda)$$

## 2.4 Solution of the first central issue in HMMs

We define,
$\alpha_i(t) = p(Q_t = i, O_1 = o_1, ..., O_t = o_t|\lambda)$
and
$\beta_i(t) = p(O_{t+1} = o_{t+1}, ..., O_T = o_T|Q_t = i, \lambda)$

### 2.4.1 Forward algorithm

1. Initialization: $\alpha_i(1) = \pi_i b_i(o_1)$
2. Recursion: $\alpha_j(t+1) = \sum_{i=1}^{N} \alpha_i(t) a_{ij} b_j(o_{t+1})$ ; $1 \leq j \leq N; 1 \leq t \leq T$
3. Termination: $p(O|\lambda) = \sum_{i=1}^{N} \alpha_i(T)$

### 2.4.2 Backward algorithm

1. Initialization: $\beta_i(T) = 1$
2. Recursion: $\beta_i(t) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_j(t+1)$ ; $1 \leq j \leq N; 1 \leq t \leq T$
3. Termination: $p(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_i(1)$

## 2.5 Solution of the second central issue in HMMs

Given the parameters of the model, the most likely sequence of states can be found using the viterbi algorithm.

#### Viterbi algorithm

1. Initialization:

$$v_j(1) = \pi_j b_j(1), \forall j \in \{1, 2, ..., N\}$$

$$bt_1(j) = 0, \forall j \in \{1, 2, ..., N\}$$

2. Recursion:

$$v_j(t) = \max_{i=1}^{N} v_i(t-1)a_{ij}b_j(o_t); 1 \leq j \leq N; 1 \leq t \leq T$$

$$bt_t(j) = \operatorname*{argmax}_{i=1}^{N} v_i(t-1)a_{ij}b_j(o_t); 1 \leq j \leq N; 1 \leq t \leq T$$

3. Termination:

$$\text{best score} = \max_{i=1}^{N} v_i(T)$$

$$\text{start of backtrace} = \operatorname*{argmax}_{i=1}^{N} v_i(T)$$

## 2.6 A few necessary definitions

1. $p(Q_t = i, O|\lambda) = \alpha_i(t)\beta_i(t)$
2. $p(Q_t = i, Q_{t+1} = j, O|\lambda) = \alpha_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1)$
3. $\gamma_i(t) = p(Q_t = i|O, \lambda) = \frac{\alpha_i(t)\beta_i(t)}{\sum_{j=1}^{N}\alpha_j(t)\beta_j(t)}$
4. $\xi_{ij}(t) = p(Q_t = i, Q_{t+1} = j|O, \lambda) = \frac{\alpha_i(t)a_{ij}b_j(o_{t+1})\beta_j(t+1)}{\sum_{j=1}^{N}\sum_{k=1}^{N}\alpha_j(t)a_{jk}b_k(o_{t+1})\beta_k(t+1)}$
5. $\sum_{t=1}^{T}\gamma_i(t)$ = Expected number of times a transistion from state i will be made to any other state
6. $\sum_{t=1}^{T}\xi_{ij}(t)$ = Expected number of times a transistion from state i to state j is made

## 2.7 Solution of the Third central issue in HMMs

Using the EM algorithm, the estimated parameters turn out to be:

$$\pi_i = P(Q_0 = i|O, \lambda') = \gamma_i(1)$$

$$a_{ij} = \frac{\sum_{t=1}^{T} p(Q_{t-1} = i, Q_t = j|O, \lambda')}{\sum_{t=1}^{T} p(Q_{t-1} = i|O, \lambda')} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)}$$

$$b_i(v_j) = \frac{\sum_{t=1}^{T} p(Q_t = i|O, \lambda')\delta(o_t, v_j)}{\sum_{t=1}^{T} p(Q_t = i|O, \lambda')} = \frac{\sum_{t=1}^{T} \gamma_i(t)\delta(o_t, v_j)}{\sum_{t=1}^{T} \gamma_i(t)}$$

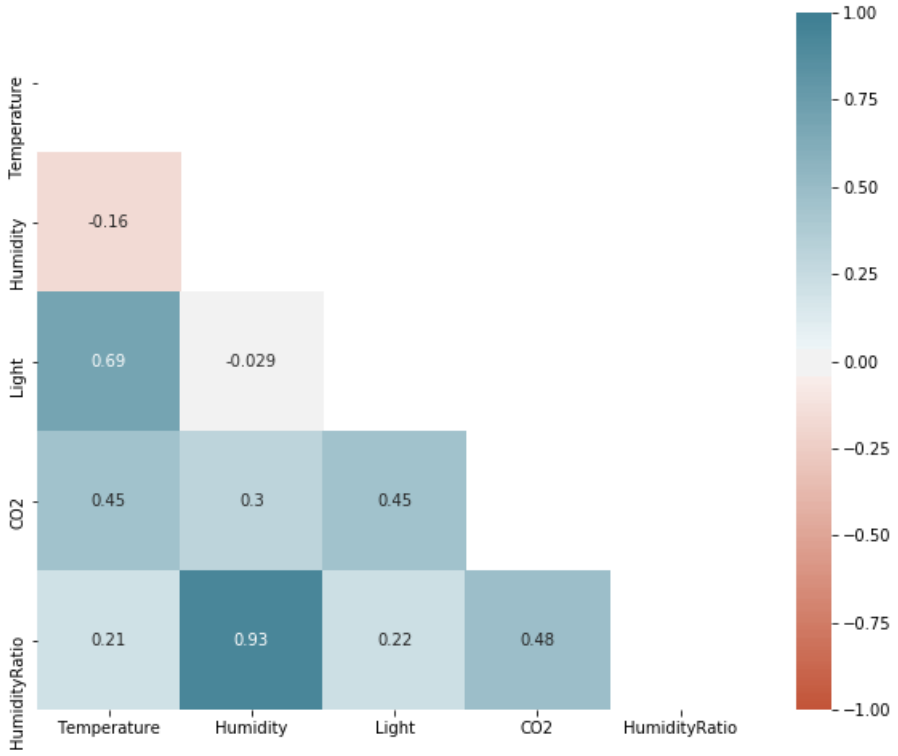
# Chapter 3

# Analysis Of The Occupancy Detection Dataset :

The Occupancy Detection Dataset [1] is a multivariate time-series dataset. The said dataset contains 7 features namely,

1. date-time in "year-month-day hour:minute:second" format
2. temperature in Celsius
3. Light in Lux
4. CO2 in ppm
5. Relative humudity in %
6. Humidity Ratio in kgwater-vapor/kg-air
7. Occupancy denoting occupancy status
   (a) 0 denotes room is not occupied
   (b) 1 denotes room is occupied

The task was to accurately predict the occupancy status of the room using features 2-to-6. Hence, featre 7 is the target feature. The correlation matrix of the features is as follows:




Since 'Relative Humidity' and 'Humidity Ratio' are highly correlated, we decided to not use one of them for prediction. Hence, we arbitrarily decide to drop the 'Relative Humidity' feature from the dataset.

From trial and error, we found that scaling the features, decreases accuracy. Hence, we have not scaled the features.

# Chapter 4

# Predictions

We applied both gaussian mixture models and gaussian hidden markov models to the occupancy detection dataset. Both statistical models gave the same accuracy of 85.21%. Moreover, the predictions made by both methods were *exactly* the same.

This means that both models extracted the same amount of information from the dataset. Also, GMMs do not have the ability to use the time-series patterns in the dataset to make predictions. Which means that Gaussian hidden markov model also was not able to use time-series patterns in the dataset to make predictions. This observation challenges the idea of treating the occupancy detection dataset as a time-series dataset.

## 4.1 A useful observation

Until now, we naively clustered the data points into 2 clusters( where the clusters represented whether the room was occupied or not). But maybe, we should cluster the datapoints into 3 clusters, where 2 clusters represent whether the room was occupied or not and the 3rd cluster is for datapoints that could not be classified into 'occupied' or 'unoccupied' status, by the model.

Hence, applying GMM on the dataset, we obtain 3 clusters, where in 2 the clusters representing occupancy status the accuracy

is 99.29%. Next, we apply GMM again on the third cluster, to divide the third cluster into 2 clusters, representing occupancy status. Doing so, yields an accuracy of 90.5%.

Overall the total accuracy of GMM(using this method) turns out to be, 97.5%, which is a over the previous accuracy of 85%.

This variant of GMMs( where you cluster the points within clusters) is called heirarchical gaussian mixture models.

# Appendices

# Appendix A

# Appendix

## A.1 Expectation-Maximization algorithm

Let, $\mathbb{X} = \{x_i\}_{i=1}^N$ and $\mathbb{Y} = \{y_i\}_{i=1}^N$

The general EM algorithm is as follows:

1. Initialize $\Theta^g$ randomly.
2. E-step: compute $Q(\Theta, \Theta^g) = E[logp(\mathbb{X}, \mathbb{Y}|\Theta)|\mathbb{X}, \Theta^g]$
3. M-step: maximize expectation computed in step 1.
   i.e.
   $$\Theta^g = \underset{\Theta}{\operatorname{argmax}}\, Q(\Theta, \Theta^g)$$
4. repeat steps 2 and 3 until convergence

# Bibliography

[1] L. M. Candanedo and V. Feldheim, "Accurate occupancy detection of an office room from light, temperature, humidity and CO 2 measurements using statistical learning models," *Energy and Buildings*, vol. 112, pp. 28–39, jan 2016.