Final report on the ICT for Health laboratories

Luca Gioacchini
s257076

A.A.2018/19

Contents

1 Parkinson's Disease

Parkinson's disease is the most common chronic neurodegenerative disorder which affects about 1.2 million people in Europe alone. The symptoms are mainly motor ones, such as tremor, rigidity, akinesia and difficulties with speech. They are due to the degeneration of the *dopamine*[1] synthesizers in the *substantia nigra*[2].
Since there is no cure, the only successful treatment is the monitoring of the patients to decide which drugs may be beneficial. An interesting analysis that can be conducted to determine if a patient is diseased or not, is the application of two Hidden Markov Machines (one for healthy people and one for diseased ones) to classify subjects' voice samples of the "a" vowel pronunciation.

The dataset is composed of ten healthy patients' voice samples and ten diseased patients' ones recorded with a sampling frequency of 8 kHz. Thanks to the FFT[3] applied to each record it is possible to estimate f_0, which is the fundamental frequency of vibration of the patient's vocal chords. The output of the FFT is a signal with peaks corresponding to f_0 and its integer multiples. After having chosen the number of samples N_s, between two signal peaks $N_s - 1$ samples are taken. Some of the new signals are used to train the Max-Lloyd algorithm in order to find N_q centroids, or representative levels. After that all the signals are quantized and the output of the quantization function is y_n.

2 Hidden Markov Machines (HMM)

By considering a discrete time random process x_n as a Markov Chain if it satisfies the *Markov properties*:

$$P(x_n|x_{n-1}, x_{n-2}, ...) = P(x_n|x_{n-1}) \tag{1}$$

it is possible to store information about the values transition probability of its values in the *Transition matrix*:

$$A = \begin{bmatrix} P(x_n = 1|x_{n-1} = 1) & P(x_n = 2|x_{n-1} = 1) & \dots & P(x_n = M|x_{n-1} = 1) \\ P(x_n = 1|x_{n-1} = 2) & P(x_n = 2|x_{n-1} = 2) & \dots & P(x_n = M|x_{n-1} = 2) \\ \vdots & \vdots & \vdots & \vdots \\ P(x_n = 1|x_{n-1} = M) & P(x_n = 2|x_{n-1} = M) & \dots & P(x_n = M|x_{n-1} = M) \end{bmatrix} \tag{2}$$

Now, by considering x_n as a state of a Finite State Machine(FSM), this is described by:

$$y_n = Q(w_n + \nu_n) \tag{3}$$

where *Q()* is a generic discrete output function which can assume K values, such as the quantization one, ν_n is a zero-mean Gaussian random variable so that ν_n and ν_{n+k} are statistically independent for any $k \neq 0$, y_n is the observable output of *Q()*, and w_n is a **hidden variable** depending on x_n. Since it has been assumed that the system state is a Markov process thanks to Equation (1), such a system is a HMM.

[1]Dopamine is a neurotransmitter responsible of memory, attention, movement, learning, ecc.
[2]Is a region of the brain, part of the basal ganglia, responsible of reward and movement
[3]Fast Fourier Transform

Besides the transition matrix, a HMM is defined also by the *Emission matrix*, which stores information about the probability of the system output given its state:

$$B = \begin{bmatrix} P(y_n = 1|x_{n-1} = 1) & P(y_n = 2|x_{n-1} = 1) & \dots & P(y_n = K|x_{n-1} = 1) \\ P(y_n = 1|x_{n-1} = 2) & P(y_n = 2|x_{n-1} = 2) & \dots & P(y_n = K|x_{n-1} = 2) \\ \vdots & \vdots & \vdots & \vdots \\ P(y_n = 1|x_{n-1} = M) & P(y_n = 2|x_{n-1} = M) & \dots & P(y_n = K|x_{n-1} = M) \end{bmatrix} \tag{4}$$

After having defined this mathematical model, thanks to the Baum-Wlech algorithm it is possible to estimate A and B from the previous ones by starting from an initial guess of them.

By considering the dataset described in Section 1, the emission matrix of the HMM is randomly initialized and it is stochastic (this means that the sum of the elements of each row is equal to 1), while the transition one is initialized in two ways: one is randomly initialized with values between 0 and 1, one is set as:

$$A = \begin{bmatrix} p & q & p & \dots & p & p \\ p & p & q & \dots & p & p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p & p & p & \dots & p & q \\ q & p & p & \dots & p & p \end{bmatrix} \tag{5}$$

where q is the probability that the state $A_{i,j}$ becomes $A_{i,j+1}$, it is equal to $(1-p)/N_s - 1$. p is set equal to 0.9.

3 Experimental Results

After having trained the two HMM, one for the healthy people and one for the diseased ones, plotting the specificity and the sensitivity can be a useful way to evaluate the machines performance. The functions output is stored in a class array c: if the *i-th* subject is classified as "healthy", the *i-th* entry of the array is 1, otherwise it is 0.

By defining the *True Positive (TP)* as the number of healthy people correctly classified and the *True Negative (TN)* as the number of diseased subjects correctly classified, the sensitivity and the specificity are determined by observing the class array:

$$TPR = \frac{TP}{\mathbb{I}(c = 1)} \tag{6}$$

$$TNR = \frac{TN}{\mathbb{I}(c = 0)} \tag{7}$$

where $\mathbb{I}(c)$ is the indicator function[4]. Equation (6) is called True Positive Rate, or*Sensitivity*, and it measures the times that the healthy subject classification is correctly performed by the first HMM on the total amount of 1s obtained. Equation (7) is called True Negative Rate, or *Specificity*, and it measures the times that the diseased patients classification is correctly performed by the second HMM on the total amount of 0s obtained.

[4]I(c=1)=1 if c=1 and 0 if c=0

3.1 Different Testing Signals

Firstly, it can be interesting to analyze the HMMs performances when different testing and training signals are used.

Testing Signals	Test. Circular		Test. Random		Train. Circular		Train. Random	
	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR
{8, 9, 10}	1	1	1	0	1	1	1	1
{1, 2, 3}	1	1	1	1	1	1	0.86	1
{4, 5, 6, 7}	0.75	1	0.75	1	1	1	1	1

Table 1: TPR and TNR values obtained with two different transition matrices and testing signals. The *Circular* label refers to the transition matrix defined in Equation (5). The Max-Lloyd algorithm has tolerance $= 10^{-3}$ and at most 200 iterations. $N_q = N_s = 8$.

Table 1 shows different values of TNR and TPR when different subsets of testing signals and two different initial transition matrices are used. The number of the signals refers to both the healthy patients ones and the diseased patients ones. By comparing the result it is clear that the circular transition matrix defined in Equation (5) leads to better performances, even if in the last entry the True Negative Rate value of the testing phase is slightly lower than the other entries because of the greater number of used signals.

3.2 Different HMMs Parameters

By recalling the definition of N_q as the number of quantization levels and N_s as the number of the states taken by the machines, it can be interesting to change these parameters in order to evaluate the performances.

HMM Param.		Test. Circular		Test. Random		Train. Circular		Train. Random	
N_s	N_q	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR
2	2	0.667	0.33	0	0.33	0.14	1	0.14	1
2	5	1	0.667	1	0.667	0.85	0.85	0.85	0.71
2	8	1	0	1	0	1	1	1	1
5	2	-	-	-	-	-	-	-	-
5	5	1	0.33	0.667	1	0.85	0.85	0.71	1
5	8	1	0.33	1	0	1	1	1	1
8	2	-	-	-	-	-	-	-	-
8	5	-	-	-	-	-	-	-	-
8	8	1	1	1	0	1	1	1	1

Table 2: TPR and TNR values obtained with different N_q and N_s. The *Circular* label refers to the transition matrix defined in Equation (5) and the training signals used are {8, 9, 10}. The Max-Lloyd algorithm has tolerance $= 10^{-3}$ and at most 200 iterations

The empty entries of Table 2 marked with a "-" are due to a software limitation: the MATLAB function `hmmdecode` works only with $N_q \geq N_s$.

By observing the obtained result it is possible to interpret the N_s parameter as an indication of how complex the HMM model is, while the N_q one indicates the precision of the data description, or approximation.

By discarding the results obtained when the transition matrix is randomly initialized (since it introduces some errors in the classification phase), it can be observed that, if one of the two parameter is fixed, the value growth of the other one leads to even better performances. This behavior can be explained by considering that high values of both parameters mean that the data are described in a very accurate way and the model can merge the voice samples and the output classes with few errors.
Obviously, when both parameters are initialized with the greater value ($N_q = N_s = 8$), the TPR and the TNR values are the greater ones.

3.3 Different Max-Lloyd Algorithm Parameters

In Table 1 and Table 2, when the signals used in the testing phase are $\{1, 2, 3\}$ and when the HMM parameters are $N_s = 8$ and $N_q = 2$, even if the classification is successfully performed, the Max-Lloyd algorithms does not converge. Because of this, can be interesting to evaluate the performances when the tolerance and the maximum number of iterations of the algorithm are modified.

Max-Lloyd Param.		Test. Circular		Test. Rand.		Train. Circular		Train. Rand.	
Tolerance	Iterations	TNR	TPR	TNR	TPR	TNR	TPR	TNR	TPR
10	10	1	0.33	1	0.67	1	1	1	1
10	500	1	0.33	1	0.67	1	1	1	1
10^{-3}	10	1	1	1	0.33	1	1	1	1
10^{-3}	500	1	1	1	0	1	1	1	1
10^{-9}	10	1	1	1	0.33	1	1	1	1
10^{-9}	500	1	1	1	0	1	1	1	1

Table 3: TPR and TNR values obtained with different tolerance and maximum iteration values used in the Max-Lloyd algorithm. The *Circular* label refers to the transition matrix defined in Equation (5) and the training signals used are $\{8, 9, 10\}$. $N_q = N_s = 8$.

Table 3 shows that when the tolerance is set as a high value, such as 10, the True Positive Rate values are not enough for a good classification, since the data approximation is not sufficient. Similar results can be obtained when the maximum number of iterations is set equal to 10, since the stopping condition is reached too early. However this behavior can be avoided by using a circularly initialized transition matrix.

4 Conclusions

According to Section 3 it is clear that the set of signals used in the training phase does not affect too much the performances. In all the measurements the circular transition matrix defined in Equation (5) is preferable and the choice of $N_s = 8$ and $N_q = 8$ is acceptable to obtain an accurate classification. Finally, the Max-Lloyd algorithm parameters should be reasonably chosen (a tolerance of 10^{-3} and at most 500 iterations shall ensure good performances).