ISSN 2007-9737

Breast Cancer Classification through Mixture of Bivariate Normal Using EM Algorithm

Gerardo Martínez-Guzmán[1], Carmen Cerón-Garnica[1,*], Jorge Alejandro Fernández-Pérez[2], Gerardo Villegas-Cerón[3]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Puebla, Mexico

[2] Benemérita Universidad Autónoma de Puebla, Instituto de Ciencias, Puebla, Mexico

[3] Benemérita Universidad Autónoma de Puebla, Facultad de Medicina, Puebla, Mexico

{gerado.martinezgu, jorge.fernandez}@correo.buap.mx, academicaceron@gmail.com, gerardo.villegasc@alumno.buap.mx

Abstract. An analysis is presented in this paper for benign and malignant diagnosis of tumors, biopsies have shown an increase of nuclear size, and changes in the texture of the tumor nucleus. In this article, an analysis is made using the unsupervised learning algorithm Expectation-Maximization (EM). Two variables are analyzed: the mean of the radius and texture of the tumors, being the former a measure of the average distances from the center of the tumor to its perimeter, and the latest is the variance of gray-scale values. Since the behavior of the said variables is similar to the mixture of normals in two opposing categories. The EM algorithms demonstrates ability to categorize the dataset into two different labels (malignant and benign). This model projects a classification with a high percentage of coincidence with the observed data.

Keywords. Maximum likelihood estimators, breast cancer, EM algorithm, Gaussian mixture model.

1 Introduction

In the descriptive analysis of the combined distribution of the variables radius_mean and texture, it is noticed that these variable cannot be described or studied by only one statistical distribution, as shown in the bar graphs.

Such situation force us to use some method of distribution combination that can describe to a large degree the data of the variables under study. This study type for these variables has not been realized using the unsupervised learning algorithm, expectation-maximization (EM), which we think largely describes the sample data for these two attributes.

One iterative method used to obtain the maximum likelihood estimation of a set of parameters in a statistical model is the EM algorithm. Initially introduced by Arthur Dempster, Nan Laird, and Donald Rubin in a 1977 publication in the Royal Statistical Society [5], this algorithm also relies on a set of unobserved parameters.

For achieving this, two types of data are considered, the observed and a set of hidden data, the unions of these form the set of complete data. The introduction of the hidden data is an artificial

Table 1. Sample data of data base

ID	Diagnosis Benign (B) Malignant (M)	radius_mean	texture
862989	B	10.49	19.29
863030	M	13.11	15.56
863031	B	11.64	18.33
863270	B	12.36	18.54
86355	M	22.27	19.67
864018	B	11.34	21.26
864033	B	9.777	16.99
86408	B	12.63	20.76
86409	B	14.26	19.65
864292	B	10.51	20.19
864496	B	8.726	15.83
864685	B	11.93	21.53
864726	B	8.95	15.76
864729	M	14.87	16.67
864877	M	15.78	22.91
865128	M	17.95	20.01

construction that, although seem surprising, favors the estimation of the parameters.

Finite mixture distributions have been applied for representing heterogeneous data, mainly because, generally it is not enough to only describe the distribution of information using an unique statistical distribution.

Hence the combination of distributions is needed to represent these kind of data. Obtaining these components lead to the estimate of proportions and parameters, where each of said components has a level of contribution to the general distribution [8, 9]. Thus, clustering the observations in different factions that share certain characteristics is required.

As one common iterative tool for maximizing the likelihood estimation of the mixture distributions [6, 7], the main aim of the EM algorithm is to use a multinomial variable that can determine the membership of each data point in the dataset to a specific group.

The main goal of this work is to classify whether a tumor presents cancer cells(malignant) or shows absence of these cells (benign). To achieve this, we consider that in previous works machine learning classification methods have been used to adjust a function that can predict the diagnosis of tumor with a new entry [4, 10, 1, 12].

However, previous works considered take into account the variables that generate the computer program that uses the curve fitting algorithm. In contrast, this algorithm computes ten features for each sample. It then calculates the mean and extreme values, along with the standard error for each feature of the image.

Finally, it returns a real-valued vector consisting of 30 variables, many of which are correlated and provide redundant information. In this work, a minimum of variables are considered with low correlation, and considering that data of these variables present a normal behavior, expectation? maximization algorithm is used for Gaussian densities that in previous works has not been analyzed and that can discriminate breast cancer with a reasonable security.

2 Data Examination

The dataset was taken from the data provided by Dr. William H. Wolberg, who works at the Hospital of the University of Wisconsin in Wisconsin, U. S. A. [12], who reviewed 569 cases of tumors in women. All records have an unique ID along with the diagnosis (malignant or benign), as well as other variables related to the tumor.

For this study, the variables considered where the mean of the radius and its texture, Breast cancer is characterized by a solid tumor, and, according to GLOBOCAN, the fourth cause of dead related to cancer in general and is the type of cancer most frequent in women and the most lethal.

Cancer is a transformative cell process and this way, cancer cells can morphologically differ from the cells that originated them [5]. On the other hand, cancer cells present genomic instability, which gives place to mutations and rearrangements in the chromatin (packaged DNA by the action of histone pro-teins), these changes

ISSN 2007-9737

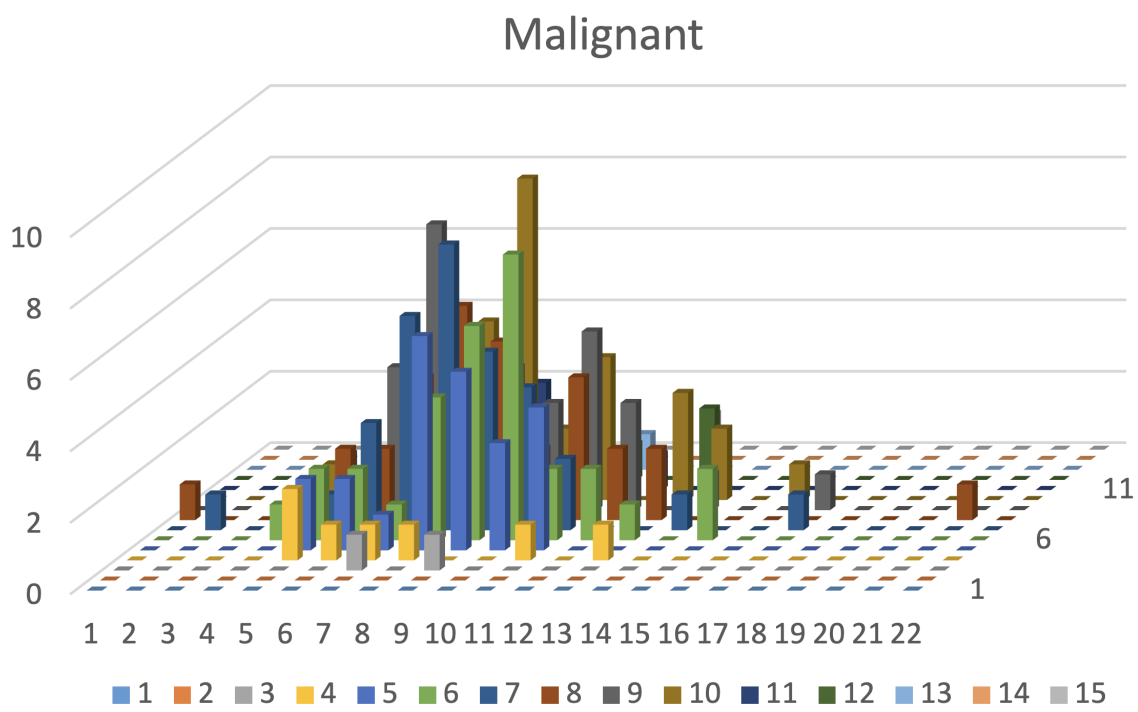

Fig. 1. An estimated probability density function of malignant tumors

allow that cancer cells develop capabilities to survive, proliferate and disseminate [3].

The analysis of the texture of the nucleus is found in the varia-tion of the intensity of the grayscale in the pixels and is employed since changes are reflected in the chromatin, allowing to determine that exists a significant difference in the values of the nuclear texture among patients with benign and malignant diagnosis. This suggests that changes in the chromatin exist in nuclear cells of malignant diag-nosis in comparison to benign. From the dataset, 357 benign are present, as well as 212 malignant cases [12]. A short example of the information available is shown in Table 1.

The following frequency histograms of the sample demonstrate an estimate of the density function form. Nevertheless, it is not the density, but, from the non-parametric perspective, it also can be interpreted as a equitable estimate.

Thus by considering the mwan of the radius and texture variables, considering that from the total observations (569), 357 indicate a benign case, meanwhile 212 cases indicate a malignant tumor, the histogram presents the distribution of a two normals mixture as shown in the following figures (1, 2).

Looking at the frequency histogram, it can be considered a behavior of mixture of normals, hence the work will be developed considering that the radius_mean and texture variables have a probability distribution completely specified, that is to say, a two normal distribution mixture. By applying the algorithm of expectation-maximization (EM), an estimate of parameters can be found, where:

$$
\begin{aligned}
1 &= [9.71, 11.12), \quad 9 = [20.99, 22.4), \quad 17 = [32.27, 33.68),\\
2 &= [11.12, 12.53), \quad 10 = [22.4, 23.81), \quad 18 = [33.68, 35.09),\\
3 &= [12.53, 13.94), \quad 11 = [23.81, 25.22), \quad 19 = [35.09, 36.5),\\
4 &= [13.94, 15.35), \quad 12 = [25.22, 26.63), \quad 20 = [36.5, 37.91),\\
5 &= [15.35, 16.76), \quad 13 = [26.63, 28.04), \quad 21 = [37.91, 39.32),\\
6 &= [16.76, 18.17), \quad 14 = [28.04, 29.45), \quad 22 = [39.32, 40.73).\\
7 &= [18.17, 19.58), \quad 15 = [29.45, 30.86),\\
8 &= [19.58, 20.99), \quad 16 = [30.86, 32.27),
\end{aligned}
$$

Once having the estimates, we will analyze to what extent the previous variables can contribute to predict whether the diagnosis confirms or denies the presence of cancer cells, by using said algorithm which can categorize these data into groups.

2.1 Expectation-maximization Algorithm

When implementing the EM algorithm, the $Y = (Y_1,\ Y_2,\ \ldots,\ Y_n)$ variable will denote a random sample of size n, where Y_i is a $p-$dimensional random vector with a density function $f(y_i)$ where $y_i \in R^p$. So $y = (y_1, y_2, ..., y_n)$ is an observed sample of Y.

Definition 1. If the density function of a random variable Y_i is:

$$f(y_i|\psi) = \sum_{k=1}^{g} \pi_k\ f_k(y_i|\theta_k),\ \ y_i \in R^p. \qquad (1)$$

It has a finite mixture distribution with g components, with a parameter vector:

$$\psi = (\pi_1,\ \ldots,\ \pi_g,\ \theta_1,\ \ldots,\ \theta_g). \qquad (2)$$

Here, $f_k(y_i|\theta_k), k = 1, 2, ..., g$, denotes the densities of the components of the mixture with parameters θ_k and weight parameters $\pi_1, ..., \pi_g$. In the most general case, it is also assumed that the functions $f_k(y_i|\theta_k)$ can belong to different parametric families.

To tun the mixture into a density function of the weight, it needs the following conditions:

$$0 \leq \pi_k \leq 1,\ \ k = 1,\ \ldots,\ g \text{ and } \sum_{k=1}^{g} \pi_k = 1. \qquad (3)$$

Note that in the last expression, one weight is defined in terms of the others, and turns redundant.

Definition 2. Let $y = (y_1,\ y_2,\ \ldots,\ y_n)$ be independent observations of a random variable, whose density function $f(y|\psi)$ is a mixture, then the function:

$$L(\psi|y) = \prod_{i=1}^{n} f(y_i|\psi) = \prod_{i=1}^{n}\sum_{k=1}^{g} \pi_k f_k(y_i|\theta_k). \qquad (4)$$

Table 2. Estimator values

θ_k	**Value of** $\hat{\theta}_k$	θ_k	**Value of** $\hat{\theta}_k$
π_1	0.72	$\rho_x y$	0.14
π_2	0.28	μ_u	17.70
μ_x	12.75	μ_v	24.54
μ_y	17.69	σ_u	3.85
σ_x	2.15	σ_v	4.21
σ_y	3.16	$\rho_u v$	-0.33

Is called maximum likelihood function. Taking the natural logarithm in $L(\psi|y)$, its $\log-$likelihood function is derived:

$$l(\psi|y) = \log L(\psi|y) = \log \prod_{i=1}^{n}\sum_{k=1}^{g} \pi_k f_k(y_i|\theta_k) \sum_{i=1}^{n} \log \sum_{k=1}^{g} \pi_k f_k(y_i|\theta_k). \qquad (5)$$

For calculating the maximum likelihood estimator $(\widehat{\psi})$, it is customary using the logarithm of the likelihood function, well, let us remember that the logarithm of the function and the function under certain regular conditions have their maximum in the same point. Hence, we must solve the likelihood equation:

$$\frac{\partial}{\partial \psi} \sum_{i=1}^{n} \log \sum_{k=1}^{g} \pi_k f_k(y_i|\theta_k) = 0. \qquad (6)$$

Owing to the existence of the logarithm of a sum, the solution of the equation is difficult. Thus, another approach that allows the maximization of the $\log$-likelihood function is required. This new procedure was first introduced by Dempster [2].

It was a mechanism to handle missing information and involves defining a new expectation that facilitates maximization, such that the parameters that maximize this "expectation" in each iteration converge to the parameters that maximize the likelihood function.

Let $y = (y_1,\ y_2,\ \ldots,\ y_n)$ be an observed sample of size n, which we will denote as the incomplete data vector, corresponding to a realization of Y, with density function $f(y|\psi)$, where ψ is the vector of parameters to be estimated. Next, consider the variable $Z = (Z_1,\ Z_2,\ \ldots,\ Z_n)$, called latent,

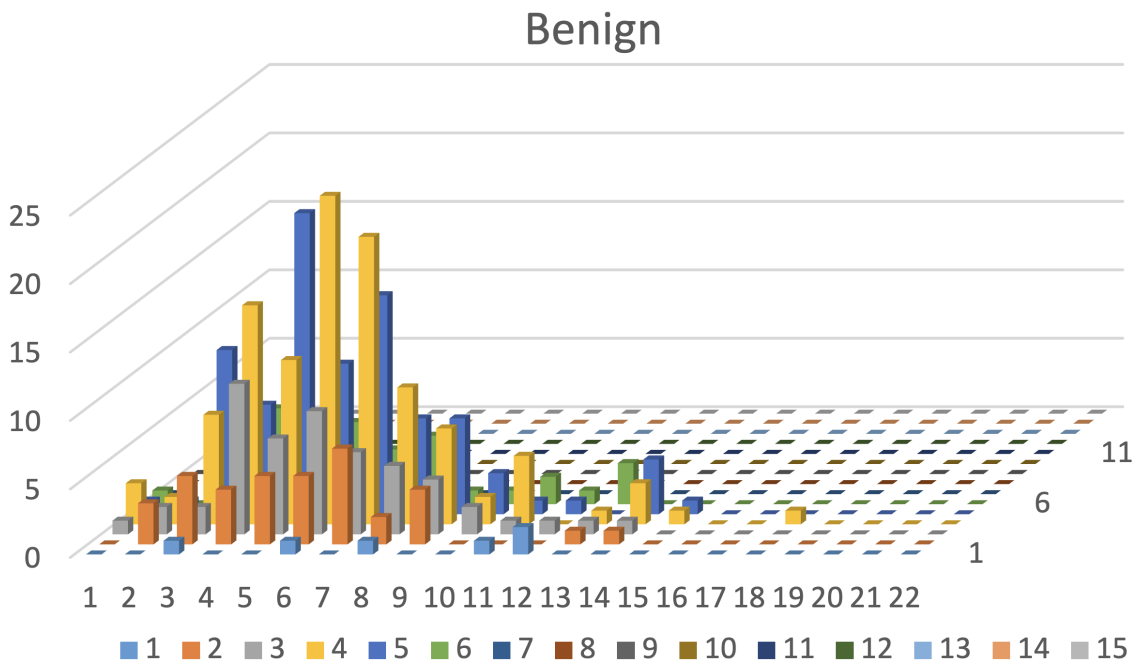

Fig. 2. An estimated probability density function distribution of benign tumors

which represents the unobserved data, and whose realization is $z = (z_1,\ z_2,\ \ldots,\ z_n)$.

Thus, the random vector $X = (Y,\ Z)$, called the complete data vector, has realization $x_1 = (y_1,\ z_1)$, $x_2 = (y_2,\ z_2)$, $\ldots$, $x_n = (y_n,\ z_n)$, such that each y_i realization always corresponds to z_i. In this regard, we assume that Z_i represents a $g-$dimensional binary indicator variable, where the $j-$th element Z_{ij} indicates the membership of the observation y_i to the $j-$th component of the mixture, with $i = 1,\ 2,\ \ldots,\ n$ and $j = 1,\ 2,\ \ldots,\ g$. Therefore, we can define Z_{ij} as:

$$Z_{ij} = z_{ij} = \begin{cases} 1 & \text{if } y_i \text{ stems from the } j-\text{th component,} \\ 0 & \text{in any other case.} \end{cases} \tag{7}$$

Due to the categorical nature of the Z_i variable, which indicates the membership of the sample points to a component (or any other part) of the mixture, the π_k weights can be interpreted as the a priori probability that the y_i observation belongs to the k-th population. This makes the assumption that Z_i follows a multinomial distribution, with just one realization across g categories and probabilities $\pi = (\pi_1,\ A\pi_2,\ \ldots,\ \pi_g)$:

$$P(Z_i = z_i) = \binom{1}{z_{i1},\ z_{i2},\ \ldots,\ z_{ig}} \pi_1^{z_{i1}}\ \pi_2^{z_{i2}}\ \ldots\ \pi_g^{z_{ig}} = \prod_{k=1}^{g} \pi_k^{z_i k}, \tag{8}$$

where:

$$\sum_{k=1}^{g} z_{ik} = 1 \quad \sum_{k=1}^{g}\sum_{i=1}^{n} z_{ik} = n. \tag{9}$$

Then:

$$P(z_{ik} = 1) = \pi_k \quad k = 1,\ 2,\ \ldots,\ g. \tag{10}$$

2.2 Gaussian mixture model

From the observation of the elections made, information about the missing data can be obtained. Hence, the density of this missing data is $h(z|y, \psi)$, which is conditioned to the perceived

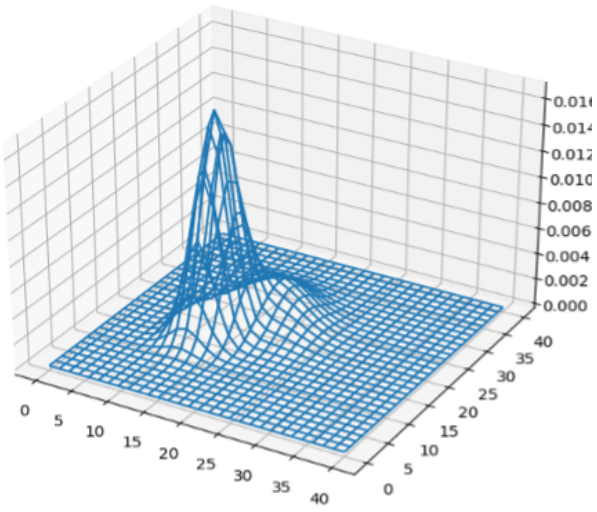

Fig. 3. Mixture of normal distribution

decisions in the sample. Therefore, by using Bayes' theorem:

$$h(z|y,\psi) = P(z_{ik}=1|Y_i=y_i) = \frac{P(z_{ik}=1)P(Y_i=y_i|z_{ik}=1)}{P(Y_i=y_i)} = \frac{\pi_k f_k(y_i|\theta_k)}{\sum_{k=1}^{g}\pi_k f_k(y_i|\theta_k)}. \tag{11}$$

Building on the previous points, we define the new expectation, or "hope," in ψ, which is linked to a likelihood function but utilizes the conditioned distribution $h(z|y,\psi)$. For the EM procedure, an initial value of the parameter ψ^0 is needed at the beginning. The algorithm then iterates, updating ψ in each step. As we observe the successive maximization of this new function, it converges to the same maximum value as the original likelihood function:

$$\varepsilon(\psi|\psi^0) = E\left[l(\psi|y,z)|Y=y,\psi^0)\right], \tag{12}$$

$$\varepsilon(\psi|\psi^0) = E\left[\sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik}\log[\pi_k\ f_k(y_i|\theta_k)]|Y=y,\ \psi^0\right] = \sum_{i=1}^{n}\sum_{k=1}^{g} E\left[z_{ik}|Y_i=y_i,\psi^0\right]\left[\log\pi_k + \log f_k(y_i|\theta_k)\right]. \tag{13}$$

However:

$$E\left[z_{ik}|Y_i=y_i,\psi^0\right] = P(z_{ik}=1|Y_i=y_i,\psi^0) = \left.\frac{f_k(Y_i=y_i|z_{ik}=1)P(z_ik=1)}{P(Y_i=y_i)}\right|_{\psi^0} = \left.\frac{\pi_k f_k(y_i|\theta_k)}{\sum_{k=1}^{g}\pi_k f_k(y_i|\theta_k)}\right|_{\psi^0} = \hat{\tau}_{ik}^{(0)}. \tag{14}$$

Therefore:

$$\varepsilon(\psi|\psi^0) = \sum_{i=1}^{n}\sum_{k=1}^{g}\hat{\tau}_{ik}^{(0)}\left[\log\pi_k + \log f_k(y_i|\theta_k)\right] = \sum_{i=1}^{n}\sum_{k=1}^{g}\hat{\tau}_{ik}^{(0)}\log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{g}\hat{\tau}_{ik}^{(0)}\log f_k(y_i|\theta_k). \tag{15}$$

After the previous calculus, the maximization of the ε function is done according to ψ. This maximization is performed in two steps because π_k appears only in the first summand, while θ_k appears only in the last summand. We begin with the maximization of the first summand that does not depend on the density functions $f_k(y_i|\theta_k)$. For that reason, we use the Lagrange multipliers:

$$\frac{\partial}{\partial_k}\left(\sum_{i=1}^{n}\sum_{k=1}^{g}\hat{\tau}_{ik}^{(0)}\log\pi_k + \lambda\left[\sum_{k=1}^{g}\pi_k - 1\right]\right) = 0, \tag{16}$$

$$\sum_{i=1}^{n}\hat{\tau}_{ik}^{(0)}\frac{1}{\pi_k} + \lambda = 0, \tag{17}$$

$$\sum_{i=1}^{n}\hat{\tau}_{ik}^{(0)} = -\lambda\pi_k. \tag{18}$$

By summing over k on both sides of the final equality, we obtain:

$$n = \sum_{i=1}^{n}\sum_{k=1}^{g}\hat{\tau}_{ik}^{(0)} = \sum_{k=1}^{g} -\lambda\pi_k = -\lambda. \tag{19}$$

Which implies that:

$$\hat{\tau}_{ik}^{(1)} = \pi_k = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_{ik}^{(0)}. \tag{20}$$

Table 3. Sample data of the distribution of mixture of normals

ID	Diagnosis Benign (B) Malignant (M)	radius_mean	texture	Mixture of Normals (1) Benign (2) Malignant
862989	B	10.49	19.29	1
863030	M	13.11	15.56	1
863031	B	11.64	18.33	1
863270	B	12.36	18.54	1
86355	M	22.27	19.67	2
864018	B	11.34	21.26	1
864033	B	9.777	16.99	1
86408	B	12.63	20.76	1
86409	B	14.26	19.65	1
864292	B	10.51	20.19	1
864496	B	8.726	15.83	1
864685	B	11.93	21.53	1
864726	B	8.95	15.76	1
864729	M	14.87	16.67	1
864877	M	15.78	22.91	2
865128	M	17.95	20.01	2
865137	B	11.41	10.82	1
86517	M	18.66	17.12	2
865423	M	24.25	20.2	2
865432	B	14.5	10.89	1
866714	B	12.19	13.29	1
8670	M	15.46	19.48	1
86730502	M	16.16	21.54	1
867387	B	15.71	13.93	1

For the maximization of the second summand with regard to θ_k, ir depends on the density function $f_k(y|\theta_k)$, and in our case, it corresponds to the Gaussian mixture:

$$f_k(y|\theta_k) = \frac{1}{(2\pi)^{\frac{p}{2}}|V_k|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(y-\mu_k)^T V_k^{-1}(y-\mu_k)\right]. \quad (21)$$

Using the estimate of μ_k, we obtain an estimate of V_k.

2.3 Start Values

The first values used for starting the algorithm, and which are also used in many other examples, are obtained by partitioning the sample in g parts, and with each, the mean of the observations is calculated: $\hat{\mu}_1^{(0)}, \hat{\mu}_2^{(0)}, ..., \hat{\mu}_g^{(0)}$.

As for the weights, they are similar following $\pi_1^{(0)} = \pi_2^{(0)} = ... = \pi_g^{(0)} = 1/g$. In [11], another form to obtain the initial values exist is presented since there are multiple ways to do it.

2.4 Halt Conditions

The ratio of difference is considered to stop the iterations:

$$\frac{\left|l(\psi^{(t+1)}|y) - l(\psi^{(t)}|y)\right|}{\left|l(\psi^{(t)}|y)\right|}. \tag{22}$$

It is utilized due to its dimensionlessness. When the maximum value of this difference is less than 10^{-6}, the algorithm stops.

3 Application of the EM Algorithm

To demonstrate that this algorithm can efficiently classify whether the tumor is benign or malignant through the attributes of mean of radius and texture, we must consider a mixture of normals with two components, each composed by two variables, radius_mean and texture. Thus the density function (21) can be defined as:

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-?_y}{?_y}\right)^2\right]\right), \tag{23}$$

where:

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}. \tag{24}$$

The database was divided into two groups, for this we calculated the mean of both variables in each register, ordered the means from smallest to largest and took the first group with the first $n_1 = 285$ registers and the second with the last $n_2 = 284$ registers, since the base consists of $n = 569$ registers. For the first group, the initial values calculated according to the average radius and texture variables:

$$\pi_1^{(0)} = \pi_2^{(0)} = \frac{1}{2}, \tag{25}$$

$$\hat{\mu}_x^{(0)} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i = 12.06, \tag{26}$$

$$\hat{\mu}_y^{(0)} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i = 16.22, \tag{27}$$

$$\hat{\sigma}_x^{(0)} = \left(\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(y_i - \hat{\mu}_x^{(0)}\right)^2\right)^{\frac{1}{2}} = 1.82, \tag{28}$$

$$\hat{\sigma}_y^{(0)} = \left(\frac{1}{n_1-1}\sum_{i=1}^{n_1}\left(y_i - \hat{\mu}_y^{(0)}\right)^2\right)^{\frac{1}{2}} = 2.36, \tag{29}$$

$$\hat{\rho}_{xy}^{(0)} = \frac{\sum_{i=1}^{n_1}\left(x_i - \hat{\mu}_x^{(0)}\right)\left(y_i - \hat{\mu}_y^{(0)}\right)}{\sum_{i=1}^{n_1}\left(x_i - \hat{\mu}_x^{(0)}\right)^2\sum_{i=1}^{n_1}\left(y_i - \hat{\mu}_y^{(0)}\right)^2} = -0.25. \tag{30}$$

For the second group, the initial values calculated based on the values of the radius_mean and texture variables were:

$$\hat{\mu}_u^{(0)} = \frac{1}{n_2}\sum_{i=1}^{n_2} u_i = 16.20, \tag{31}$$

$$\hat{\mu}_v^{(0)} = \frac{1}{n_2}\sum_{i=1}^{n_2} v_i = 22.37, \tag{32}$$

$$\hat{\sigma}_u^{(0)} = \left(\frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(u_i - \hat{\mu}_u^{(0)}\right)^2\right)^{\frac{1}{2}} = 3.56, \tag{33}$$

$$\hat{\sigma}_v^{(0)} = \left(\frac{1}{n_2-1}\sum_{i=1}^{n_2}\left(v_i - \hat{\mu}_v^{(0)}\right)^2\right)^{\frac{1}{2}} = 3.53, \tag{34}$$

$$\hat{\rho}_{xy}^{(0)} = \frac{\sum_{i=1}^{n_2}\left(u_i - \hat{\mu}_u^{(0)}\right)\left(v_i - \hat{\mu}_v^{(0)}\right)}{\sum_{i=1}^{n_2}\left(u_i - \hat{\mu}_u^{(0)}\right)^2\sum_{i=1}^{n_1}\left(v_i - \hat{\mu}_v^{(0)}\right)^2} = -0.15. \tag{35}$$

The algorithm starts to iterate and stops in the iteration where it is fulfilled (see Table 2):

$$\frac{\left|l(\psi^{(t+1)}|y) - l(\psi^{(t)}|y)\right|}{|l(\psi^{(t)}|y)|} < 10^{-6}. \tag{36}$$

Table 4. Hypothesis for tests

	Hypothesis H_1: Malignant Tumor	**Hypothesis H_0: Benign Tumor**
Positive in test	True positive (TP), $H_1\|H_1$	False positive (FP), $H_1\|H_0$
Negative in test	False negative (FN), $H_0\|H_1$	True negative (TN), $H_0\|H_0$

Table 5. Distribution of hypothesis

	TP	**FP**	**FN**	**TN**
Patients	127	27	85	330
Percentage	22.32%	4.75%	14.94%	58.00%

With these values, we obtain the distribution of the mixture of normals, some results can be seen in Table 3. As seen, there are differences in the results, for example, in the registers 2, 14, 22 and 23, the model classifies them in component one (benign), but the diagnosis is labeled as malignant.

Figure 3 shows the distribution of the mixture of normals. Utilizing the results of the EM algorithm to perform hard clustering of the observations, and to analyze the produced errors, Table 4 and 5 are obtained. We represent P_{TP} as the probability that an individual with a malignant tumor has a positive result:

$$P_{\mathrm{TP}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = P(H_1|H_1), \tag{37}$$

$$P_{\mathrm{TP}} = \frac{127}{212} = 0.5991. \tag{38}$$

And we represent P_{FP} as the probability that a healthy individual has a positive result:

$$P_{\mathrm{FP}} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}} = P(H_1|H_0), \tag{39}$$

$$P_{\mathrm{FP}} = \frac{330}{357} = 0.9244. \tag{40}$$

For this particular study, patients with malignant tumor are detected in approximately 60%, however, patients with benign tumor are detected in 92%.

4 Conclusions

We can observe that, from a total of 569 observations, 62.7%, 357 cases, show the absence of cancer cells, that is benign tumors. On the other hand, 212 cases are malignant, that is to say, 37.3% manifest cancer cells. Using the model made with the mixture of normals, a coincidence of about 80% of the total cases was found having a mayor difference of coincidences in the malignant tumors.

The results obtained from the mixture of normals reveal a coincidence of approximately 92% of the total benign cases. On the contrary, for the malignant cases the coincidence is about 60% of the total of cases for malignant tumors. We can conclude that the presented model has a good acceptance for benign tumors, however, for malignant cases the prediction is not as satisfying.

In this last case, we can argue that the data provided by Dr. Wolberg does not represent a typical distribution of medical analysis. Moreover, the distribution of benign versus malignant cases is unbalanced, with more benign tumors than malignant ones. Thus that could explain why there is a higher coincidence with benign cases than for malignant tumors.

References

1. **Agarap, A. F. (2018).** On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, pp. 5–9. DOI: 10.1145/3184066.3184080.

2. **Dempster, A. P., Laird, N. M., Rubin, D. B. (1977).** Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (Methodological), Vol. 39, No. 1, pp. 1–38.

3. **Denais, C., Lammerding, J. (2014).** Nuclear mechanics in cancer. Cancer Biology and the Nuclear Envelope, pp. 435–470. DOI: 10.1007/978-1-4899-8032-8_20.

4. **Gharagyozyan, H. (2019).** A practical application of machine learning in medicine. www.macadamian.com/learn/a-practical-application-of-machine-learning-in-medicine/.

5. **Hanahan, D., Weinberg, R. (2011).** Hallmarks of cancer: The next generation. Cell, Vol. 144, No. 5, pp. 646–674. DOI: 10.1016/j.cell.2011.02.013.

6. **Levine, R. A., Casella, G. (2001).** Implementation of the Monte Carlo EM algorithm. Journal of Computational and Graphical Statistics, Vol. 10, No. 3, pp. 422–439.

7. **McCulloch, C. E. (1998).** Review of the EM algorithm and its extensions. Journal of the American Statistical Association, Vol. 93, No. 441, pp. 403–404.

8. **McLachlan, G., Peel, D. (2000).** Finite mixture models. Wiley Series in Probability and Statistics. DOI: 10.1002/0471721182.

9. **Mergersen, K. L., Robert, C., Titterington, M. (2011).** Mixtures: Estimation and applications. Wiley. DOI: 10.1002/9781119995678.

10. **Murphy, A. (2021).** Breast cancer wisconsin (diagnostic) data analysis using GFS-TSK. Explainable AI and Other Applications of Fuzzy Techniques, pp. 302–308. DOI: 10.1007/978-3-030-82099-2_27.

11. **Redner, R. A., Walker, H. F. (1984).** Mixture densities, maximum likelihood and the EM algorithm. SIAM Review, Vol. 26, No. 2, pp. 195–239.

12. **Wolberg, W., Mangasarian, O., Street, N., Street, W. (2023).** Breast cancer wisconsin (Diagnostic). archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.

Article received on 17/04/2024; accepted on 28/06/2024.
**Corresponding author is Carmen Cerón-Garnica.*