

Selecting Features in Dichotomous Classification Problems

Leonardo Araujo

Abstract—In the present paper we propose the usage of logistic regression which is applied in three examples of dichotomous classification (heart disease, breast cancer and chemotherapy response). We present the results for models of different orders, where the features to be used are selected according to a deviance and/or hit-rate criteria. It is here also presented the results for the dimensionally reduced version of the problem when multidimensional scaling is applied and the principal components are used as features to predict the model's output. This approach has shown a considerable importance when dealing with problems with a very large dimensionality, since it is capable of providing a slightly better result with a tremendous reduction in computational time.

Index Terms—feature selection, bernoulli trial, likelihood, maximum likelihood estimator, logistic regression, multidimensional scaling, heart disease, breast cancer, chemotherapy sensitivity.

I. INTRODUCTION

The aim of this paper is to present the method of logistic regression applied to binary classification in problems where the output are dichotomous variables. In order to understand the logistic regression, there is a previous overview on statistical topics related: bernoulli distribution, bernoulli trial, likelihood, maximum likelihood estimator and likelihood ratio test. Then the logistic regression method is presented itself and afterwards its application to three different classification problems: heart disease, breast cancer and chemotherapy sensitivity response.

In classification problems, the large number of features might be an obstacle to deal with the problem at hand and it might also be prejudicial to the classifying system performance, because some features might frustrate the system performance. It is an important task then to select the relevant features and discard the disguising ones. We present here three feature selection procedure: the first one is based solemnly on the deviance of the model when that feature at hand it added to the model; the second approach is based solemnly on the hit-rate performance when the feature is added to the model; and the third approach consists on selecting the best hit-rate subject to keep the deviance of model always decreasing as new features are added.

Maybe the features are not adequate to be used in the classifications problem, maybe it would be necessary to extract further information before using them in a classification task. In many situations it is also common to find data which are highly dimensional. There could be so many features that it gets hard to decide what are the relevant ones. We propose here the usage of a multidimensional scaling procedure to find a representation of the data in another space, respecting the dissimilarity between the samples. This new representation is such that the dimensions are chosen so the most of the variance of the signal is represented in as few components as possible. The dimensions (principal components) are ordered according to how much variation of the original data they are able to represent.

II. BERNOULLI DISTRIBUTION

The Bernoulli distribution, named after Swiss scientist Jacob Bernoulli, is a discrete probability distribution, which takes only two values: 1 with probability p (success) and 0 with probability $q = 1 - p$ (failure). For a random variable X with a Bernoulli distribution,

$$P(X = 1) = 1 - P(X = 0) = 1 - q = p. \quad (1)$$

The probability mass function for this distribution is given by

$$f(x; p) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

That may also be expressed as

$$f(x; p) = p^x(1 - p)^{1-x} \text{ for } x \in \{0, 1\}. \quad (3)$$

The expected value and the variance of a random variable X with such distribution are easily calculated: $E(X) = p$ and $\text{var}(X) = p(1 - p)$.

III. BINOMIAL DISTRIBUTION

Taking a series of n independent binary outcome experiments (Bernoulli trial), the number of successes in a sequence is distributed according to a discrete probability distribution known as binomial distribution. For the simple case when $n = 1$, the binomial distribution is a Bernoulli distribution.

If each experiment on a bernoulli trial has a chance p of happening, the change of it happening k times in a series of n trials (and so not happening $n - k$ times) is given by the probability of a random variable X with a Binomial distribution to be equals to k (see figure 1). X follows a binomial distribution with parameters n and p ; and we write it as $X \sim B(n, p)$. The probability of taking k success out of n trials is given by

$$f(k; n, p) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}. \quad (4)$$

The cumulative distribution function (see figure 2) is given, for $k \in \mathbb{N}$, by the following expression:

$$F(k; n, p) = P(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i (1 - p)^{n-i}. \quad (5)$$

The mean and variance of a binomial distributed random variable $X \sim B(n, p)$ are given: $E[X] = np$ and $\text{var}(X) = np(1 - p)$.

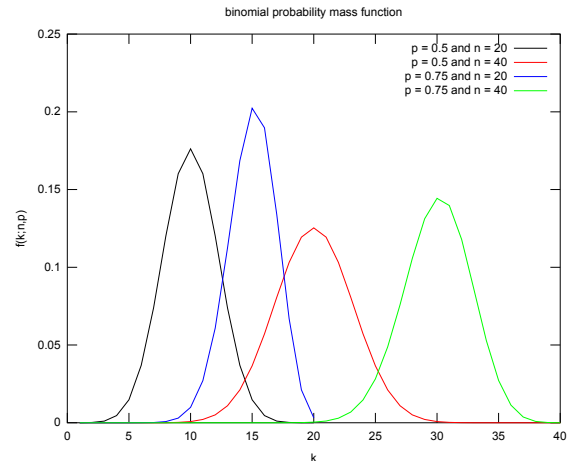


Fig. 1: Binomial probability mass function.

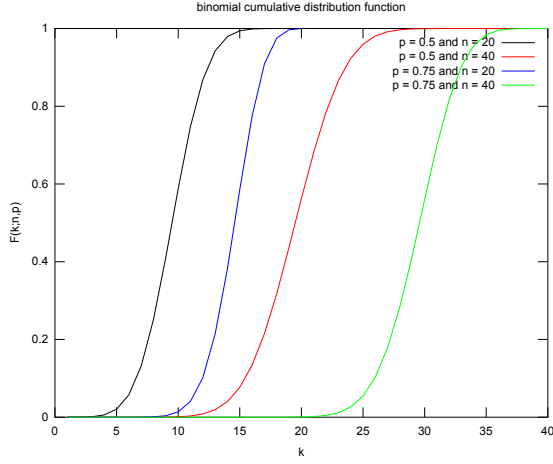


Fig. 2: Binomial cumulative distribution function.

IV. LIKELIHOOD

Likelihood plays an important role in statistical inference, it is a function of the parameters of a statistical model. It may be seen as a reverse version of conditional probability, what from the Bayes' theorem is expressed as:

$$L(B|A) = P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (6)$$

The conditional probability of the parameters B given the observation A is the likelihood function, that is calculated through Bayes using the priors probabilities and the posterior probability of the observation given the parameters of the statistical model.

The aim of the likelihood function is to find the parameters of a statistical model that make the observations most likely, due to its definition as the conditional probability of the parameters given the observations. As a more formal definition, likelihood function is conditional probability function considered as a function of its second argument (A), when the first is held fixed ($B = b$). Any function proportional to this one is also a likelihood function.

$$L(b|A) = \alpha P(A|B = b) . \quad (7)$$

The constant of proportionality is given by $\alpha = P(B = b)/P(A) > 0$. The numerical value of $L(b|A)$ alone is immaterial; what matters is the relations to other likelihoods $L(b_1|A)/L(b_2|A)$, in which case the constant of proportionality is irrelevant.

A. Likelihood of a Bernoulli Trial Experiment

Consider the simple example of a Bernoulli trial experiment, which we know the outcome and we want to determine the most likely parameter \hat{p}^* that leads to this observation. For this experiment we know exactly the number of trials n observed and the number of k of a certain outcome. So, the likelihood of a parameter $p = \hat{p}$ given the data (n and k) is

$$L(p = \hat{p}|data) = \binom{n}{k} p^k (1-p)^{n-k} . \quad (8)$$

The maximum likelihood is found taking the derivative of the likelihood function in relation to p and finding the value where it is equals to zero.

$$\frac{dL}{dp} = p^{k-1} (1-p)^{n-k-1} (k - np) . \quad (9)$$

Making $dL/dp = 0$ leads to $\hat{p}^* = k/n$, which is the maximum likelihood estimator for the parameter p of the statistical model.

V. MAXIMUM LIKELIHOOD ESTIMATION

Maximum likelihood estimation (MLE) is a statistical approach to estimate one model's parameters, like the well known least squares (LS) method (multiple linear regression model). Maximum likelihood and least squares are different approaches that happen to give the same result when the measurement errors are independent and normally distributed with constant standard deviation.

Sometimes we can write the likelihood equation as a function of the parameters; then we may differentiate it and find the parameters which satisfy the gradient equals to zero, which is the peak of the likelihood function. It represents the maximum likelihood estimate of the parameters. On the other hand, often it is not possible to analytically write the gradient of the likelihood function, or it may be too complicated and we may choose to calculate it numerically. It might be computationally expensive to calculate the likelihood for thousands of possible parameters values evaluated at small steps, in order to achieve a certain minimum tolerance by which you are happy for your estimates to be out. The computational cost may increase even further when the number of parameters grows. For this reason, the use of optimization techniques is indispensable in the numerical calculation of the maximum likelihood parameters estimate.

For computational reasons, it is preferable to use the logarithm of the likelihood instead of the likelihood itself, since the likelihood usually assumes small values, and after multiplying many such small numbers, the result might be too small to be represented digitally. This situation occurs very often, since to calculate likelihoods when we are often multiplying the probabilities of lots of rare but independent events. Using the log-likelihood, we will just have to add numbers instead of multiply.

The process of model identification through maximum likelihood estimation might have more than one solution, what typically may occur when there are many parameters being estimated and the data is not capable of uniquely defining a set of parameters. When it happens, we say that such a model is under-identified.

Another usual practical problem are the local minima. The log-likelihood surface may have several local minimum points, and the optimization technique might get stuck at these points, returning a sub-optimal solution. To avoid this kind of problem, it is a good practice to make good models specifications, choose an appropriate optimization algorithm and sensible starting values.

To use the maximum likelihood estimate, we need at first to the joint probability of taking all observations. If we may consider that the observations are independent one from another and identically distributed (i.i.d.), we will have the following

$$P(x_1, x_2, \dots, x_n | \theta) = P(x_1 | \theta) P(x_2 | \theta) \dots P(x_n | \theta) . \quad (10)$$

So, given a set of observations x_1, \dots, x_n , the likelihood of the parameter θ is given by

$$L(\theta | x_1, \dots, x_n) = P(x_1, \dots, x_n | \theta) = \prod_{i=1}^n P(x_i | \theta) . \quad (11)$$

The log-likelihood is then expressed as

$$\log L(\theta | x_1, \dots, x_n) = \sum_{i=1}^n \log P(x_i | \theta) \quad (12)$$

The method of maximum likelihood estimation consists in finding the parameter $\hat{\theta}$ which maximizes the expression above.

$$\hat{\theta}_{mle} = \arg \max_{\theta \in \Theta} \log L(\theta | x_1, \dots, x_n) . \quad (13)$$

VI. LIKELIHOOD RATIO TEST

The likelihood ratio test is the mean to compare the data likelihood against two different hypothesis (the alternate hypothesis (H_A) and the null hypothesis (H_0), a more restricted one we usually want to falsify). The likelihood ratio test answers the question: which hypothesis is the most likely to give rise to the observed data?

To perform this test is to calculate the statistical deviance between the null hypothesis and the alternate hypothesis:

$$D = -2 [\log L(H_0) - \log L(H_A)] . \quad (14)$$

The value achieved for the deviance is used to accept or reject the null hypothesis. If the value of the deviance is too small (what depends on the significance level of the test, i.e., on what probability of rejection of a null hypothesis that is true is considered tolerable), the null hypothesis is rejected. The rejection of the null hypothesis is justified by the Neyman-Pearson lemma.

The difference between the two likelihoods is multiplied by the factor 2 only for technical reasons, so that this quantity will be distributed as a chi-square distribution. The statistical significance of the results may be drawn from the standard chi-square significance levels. A chi-square probability of 0.05 is commonly used as a threshold value to reject the null hypothesis, so if we get a deviance result of 0.05 or less, we may reject the null hypothesis and say that the difference we see is likely due to some factor other than chance.

Consider the situation where we want to know if a certain coin is biased or not. We toss it 100 times and the maximum likelihood estimate give us a result $p = 0.56$. We state then the null hypothesis as: the coin is fair ($p = 0.5$); and the alternate hypothesis as: the coin is not fair ($p \neq 0.5$). The likelihood ratio test give us $D = -2(-3.247 + 2.524) = 1.446$. We conclude then that we have no reason to reject the null hypothesis that the coin is fair. The data are indeed consistent with the coin being a fair coin.

When considering the case of model fitting, the deviance may be written in the following form:

$$D(y) = -2 \left[\log p(y|\hat{\theta}_0) - \log p(y|\hat{\theta}_A) \right] , \quad (15)$$

where the parameters $\hat{\theta}_0$ are the fitted values for a model M_0 , what represents a restricted model version, and so relates to the null hypothesis. The parameters $\hat{\theta}_A$ are the parameters for a ‘full model’, a model where there are as many parameters as observations, and so the model fits exactly to the data, although it is a more complex model version. In this situation, the deviance (also known as residual deviance) is used to assess the fit of the overall model. The smaller the deviance the better the fit of the model. The deviance can be compared to a chi-square distribution, which approximates the distribution of the deviance.

VII. LOGISTIC REGRESSION

Consider the case in which we want to predict a dichotomous outcome Y based on set of independent variables X_1, \dots, X_N . The choice for the X s is flexible, we may chose exposure variables, control variables or even combinations of such variables of interest (the product of two of them, for example). Logistic regression is a mathematical modeling approach to solve the multivariable problem of finding a relation that describes the relationship of several X s and a dichotomous dependent variable Y .

Considering an event, whatever event it is, its probability of occurrence is p , a number between 0 and 1. the odds in favor of this event is the quantity $\frac{p}{1-p}$, and the odds against it is $\frac{1-p}{p}$. For example, the odds of taking a certain number on a fair dice is $\frac{1/6}{1-1/6} = \frac{1}{5}$, but not $\frac{1}{6}$, and it is often said, for this example, that the odds are 1 : 5 (one to five).

The *logit* of a probability p (a number between 0 and 1), is the logarithm of the odds

$$\text{logit}(p) = \log \left(\frac{p}{1-p} \right) = \log(p) - \log(1-p). \quad (16)$$

If we have two probabilities p_1 and p_2 , the logarithm of the odds ratio R is the difference of the logit of the two probabilities

$$\begin{aligned} \log(R) &= \log \left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)} \right) \\ &= \log \left(\frac{p_1}{1-p_1} \right) - \log \left(\frac{p_2}{1-p_2} \right) \\ &= \text{logit}(p_1) - \text{logit}(p_2) . \end{aligned} \quad (17)$$

The logistic function is given by the inverse-logit. Considering $z = \text{logit}(p)$, the inverse-logit of z is given

$$\text{logit}^{-1}(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} = p. \quad (18)$$

The logit model function was invented to describe population growth. It was introduced and coined by the Belgian mathematician Verhulst in the 19th century (Cramer, 2003). The logistic function was originally proposed to describe population growth over time and it was a solution to the differential equation of population growth where it was assumed that the growth rate is proportional to the population size with an extra term to represent the increasing resistance to the growth of large populations (a saturation term).

The logistic regression consists of fitting data to a logistic function (a sigmoid curve), which is used to predict the probability of occurrence of an event. Given a input data $\mathbf{X} = [X_1, \dots, X_n]^T$, the logistic function gives the probability of a certain event occurring for these input. The logistic function has the form (see figure 3):

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}} , \text{ where} \quad (19)$$

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n , \quad (20)$$

β_0 is known as intercept, the point where the curve intercepts the y-axis, and β_1, \dots, β_n are the regression coefficients of the independent variables X_1, \dots, X_n , respectively. In essence, then, z is an index that combines the X s. Each regression coefficient describes the contribution of its independent variable to the probability of the outcome. Logistic regression is then a useful manner of describing the relationship of the independent variables to a binary event in terms of the probability of occurrence of that event.

Logistic functions may be also used in statistics as the cumulative distribution function of the logistic family of distributions. Since it is now our aim here, it will not be further described here. It is also used in many other applications, from neural networks (it is a common choice for the activation functions) to medicine (modeling of growth of tumors) or language (used to model language change).

The logistic function $f(z)$ ranges between 0 and 1 in a monotonic increasing way (s-shaped curve), what makes it a good candidate when a probability is to be estimated.

We might wonder what is the meaning of the logit function and what is the relation to the odds ratio. Using the logistic model into the logit function, we get:

$$\begin{aligned} \text{logit}(p) &= \log \left(\frac{p}{1-p} \right) \\ &= \log \left(\frac{\frac{1}{1+e^{-z}}}{1 - \frac{1}{1+e^{-z}}} \right) \\ &= \log \left(\frac{1}{e^{-z}} \right) = \log(e^z) = z \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n . \end{aligned} \quad (21)$$

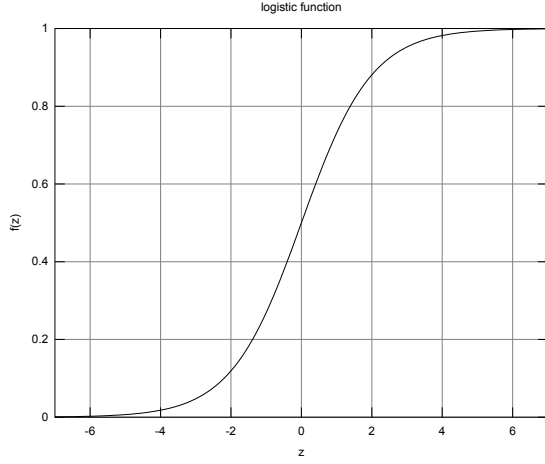


Fig. 3: Logistic Function

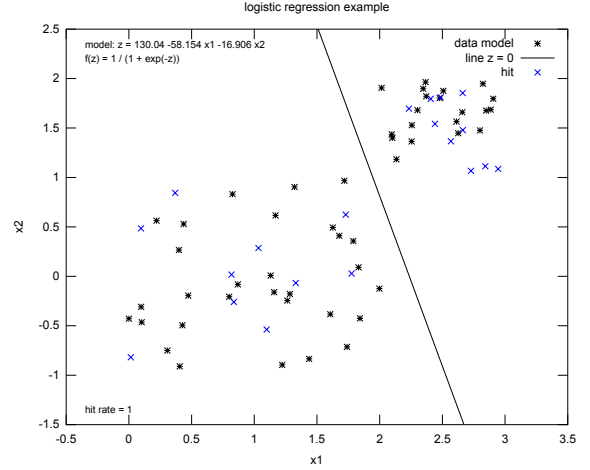


Fig. 4: Logistic Model example.

The logit function may be seen as a linear combination of the independent variables plus a constant, what defines a hyperplane in the input space. Given the logistic model, the odds may be calculated from it:

$$\text{odds} = \frac{p}{1-p} = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n). \quad (22)$$

The probability of a certain outcome y may be expressed as $P(Y = y|X_1, \dots, X_n)$. Using the logistic model defined above, this probability is expressed then as

$$P(Y = y|X_1, \dots, X_n) = \frac{1}{1 + e^{-z}}. \quad (23)$$

We could use p or $P(Y = y|X_1, \dots, X_n)$ interchangeably, but the last notation emphasizes that this probability is determined by a logistic model involving independent variables. The parameters β_1, \dots, β_n are still unknown and we need to estimate them based on the data obtained on the X s and Y for a group of samples. Once those parameters are determined, we may use the model to predict the probability of an outcome (dependent variable) based on the input data (independent variables). The parameters estimators $\hat{\beta}_1, \dots, \hat{\beta}_n$ are obtained using a maximum likelihood method.

A. Example

Here, we present a simple example of a logistic regression used in the classification of two dimensional data. Two classes of random normally distributed data with unitary standard deviation are generated. The first one has mean $[2 \ 1]^T$ and the second has mean $[0 \ -1]^T$. It was generated 20 points of the first class and 30 of the second. These data are used to create a logistic model. After the model was established, data is generated according to the distributions mentioned before and they are classified by a logistic model. After, we shall verify if the classification was successful or not. The logistic model estimated is defined by the parameters in table I and illustrated in figure 4.

parameter	value
β_0	130.04
β_1	-58.154
β_2	-16.906

TABLE I: Model's parameters value for this example.

VIII. MULTIDIMENSIONAL SCALING

“Multidimensional scaling, then, refers to a class of techniques. These techniques use proximities among any kind of objects as input. A proximity is a number which indicates how similar or how different two objects are, or are perceived to be, or any measure of this kind. The chief output is a spatial representation, consisting of a geometric configuration of points, as on a map. Each point in the configuration corresponds to one of the objects. This configuration reflects the “hidden structure” in the data, and often makes the data much easier to comprehend.” (Kruskal and Wish, 1978)

The input for a MDS method is a dissimilarity (or similarity) matrix Δ , as below:

$$\Delta = \begin{pmatrix} \delta_{1,1} & \delta_{1,2} & \dots & \delta_{1,N} \\ \delta_{2,1} & \delta_{2,2} & \dots & \delta_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{N,1} & \delta_{N,2} & \dots & \delta_{N,N} \end{pmatrix} \quad (24)$$

The dissimilarity (distance) or data value connecting object i with object j we represent by $\delta_{i,j}$.¹

The output of a MDS method is a set of N R-dimensional vectors representing the objects (or stimulus) subjected to the current study:

$$\begin{aligned} \mathbf{x}_1 &= (x_{1,1}, \dots, x_{1,R})^T \\ \mathbf{x}_2 &= (x_{2,1}, \dots, x_{2,R})^T \\ &\vdots \\ \mathbf{x}_N &= (x_{N,1}, \dots, x_{N,R})^T \end{aligned} \quad (25)$$

To calculate the MDS, we should start by calculating the dissimilarity matrix \mathbf{D} (where, in this case, $\mathbf{D} \approx \Delta$) of distance between samples of our database. We build then a matrix \mathbf{A} such that: $[\mathbf{A}]_{i,j} = a_{i,j} = -\frac{1}{2}d_{i,j}^2$. A matrix $\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}$ is calculated, where \mathbf{A} is given previously and \mathbf{H} is the following matrix $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^T$. The matrix \mathbf{B} is usually a positive-semidefinite matrix, so the singular value decomposition (SVD) applies: $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. \mathbf{B} is a matrix of rank p , then $n - p$ eigenvalues of \mathbf{B} are null and we have $\mathbf{B} = \mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1^T$, where $\mathbf{\Lambda}_1 = \text{diag}(\lambda_1, \dots, \lambda_p)$ and $\mathbf{V}_1 = [v_1, \dots, v_p]$. As $\mathbf{B} = \mathbf{X}\mathbf{X}^T$, \mathbf{X} is given by $\mathbf{X} = \mathbf{V}_1\sqrt{\mathbf{\Lambda}_1}$, and we have then calculated the matrix \mathbf{X} , which are the representation

¹In many situations there may be no effective difference in meaning between $\delta_{i,j}$ and $\delta_{j,i}$, and there may be no meaning at all for $\delta_{i,i}$, so that the data values may not form an entire matrix, but only part of one.

of the original data in a new vector space where each dimension is a principal component.

IX. APPLICATIONS

In this section we present three examples of application of the logistic function to create models for pattern classification from data sets. The methodology used in all three of them is the same and is explained below. The data sets used may be found in the Machine Learning Repository (UCI, 2010) (for the first two examples) or in Anderson Cancer Center, Houston-Texas (Hess et al., 2006).

The data used consists of a table with a certain number of independent variables (that might be real numbers or categorical, represented by binary numbers) and one output dependent variable (which is always categorical and represented by a binary number). The problem proposed is to use the independent variables to infer the value of the dependent categorical variable. In the three examples presented below, the biological data collected is used to infer the presence or not of a disease (heart disease and breast cancer) or the response to a treatment (cancer chemotherapy).

A. Methodology

The outcome of our data, the dependent variable, may be considered as a Bernoulli random variable with parameter p , which reflects the probability of presence of the disease. We want to impose a model where the independent variables lead to the dependent variable. We adopt here the logistic model, considering the independent variables as the variables x s, which are weighted according to some β s parameters, leading to z , to which we finally apply the inverse-logit, resulting in a value that is further rounded, so that we get a value either 0 or 1, and this is assumed as our dependent variable predicted by the logistic model.

To each data set, we propose a logistic model where the features are inserted one per iteration according to their deviance performance. As the complexity of the model increases with more features added to the logistic model, the deviance value decreases, as might be observed in the examples below. In both examples, both data sets are divided into a training data set (which will be used to derive the model) and a test data set (which is used to test the model's performance). At each iteration a new logistic model is obtained and tested, resulting in a hit rate value. The figure 7, 8 and 9, in the examples below, shows the progress of the deviance and hit-rate through the iterations of the method. Along the x-axis are displayed the id of each feature that is added to the model in each iteration. For a certain iteration, the features comprised by the model are the feature just added in this iteration, which is displayed on the x-axis, and all features added in previous iterations. The first y-axis shows the deviance of the model and the second y-axis the hit rate.

In order to create a model and test it, the data set was divided into two: 80% of the data was used to create the model based on logistic regression; and the other 20% was used to inquire the system performance and so obtain the system hit-rate. The smallest dataset (cancer chemotherapy response) here used has 82 samples, what leads to 66 samples to build the model and 16 samples to test its performance. The data sets for the heart disease and breast cancer have considerably more samples, 270 and 569 respectively.

At each iteration the features are chosen to be added to the model according to three criteria. First criteria: it is chosen the feature that leads to the minimum deviance model. The results from this approach are displayed in figures 7a, 7b, 9a, 8a, 8b and 9b. The second criteria used consist on selecting the feature that leads to the minimum hit-rate with the restriction that the deviance may not be bigger than the deviance in the previous iteration. The figures illustrating the results

achieved through this approach are shown in figures 7e, 7f, 9e, 8e, 8f and 9f. The third and last criteria used is regarding only the hit-rate. So it was selected the feature that leads to the minimum hit-rate. The figures illustrating the results using this criteria are in figures 7c, 7d, 9c, 8c, 8d and 9d.

When the data used is highly dimensional and there are not so many samples to fill in properly the samples space, it might be useful to use a dimensionality reduction technique. We propose here the usage of Multidimensional Scaling (MDS), where the mere distance between samples is used to create a representation in another space in which the dimensions are arranged so that most of the variance of the data is comprised by the least number of dimensions as possible. The figures 8a, 8b, 8e, 8f, 8c and 8d show the application of such procedure. The x-axis, instead of displaying the original data features, shows the principal components (PCs) that are added to the logistic model. The MDS procedure orders the PCs according to the variance of the data represented by them. It might be observed that the features were not selected according to the order of variance.

With a simple analysis of the figure 6, we can observe that the distance between samples in the heart disease and breast cancer problems are, in general, smaller than the distances displayed in the PCR problem. As the number of dimensions involved in the PCR problem is many times greater than dimensions on the other two problems, it is natural to expect to find bigger inter spaces between samples, mainly when the set of samples is small, as is the situation on the PCR problem, where we clearly see the effects of the curse of dimensionality (also known as the Hughes effect). When we perform the MDS, the samples in the lower dimensional problems might easily be well represented with only two PCs, as is the case of the heart disease and breast cancer problems (see figures 6d and 6e). When the curse of dimensionality plays its role, the samples might not be well represented by only two PCs, as seen in the PCR problem (see figure 6f).

B. Heart Disease

The first is the Heart Disease Data Set. It has a total of 270 samples with 13 features as described in the table II below. The aim of the problem is to classify the samples, based on their features, as 'with heart disease' or 'without heart disease'.

id	feature
1	age
2	sex
3	chest pain type
4	resting blood pressure
5	serum cholestoral
6	fasting blood sugar
7	resting electrocardiographic results
8	maximum heart rate
9	exercise induced angina
10	oldpeak
11	slope of the peak
12	number of major vessels
13	normal, fixed, reversible

TABLE II: Data Set Features.

The figure 7a shows the performance of the model as new features are added to it, according to a minimum deviance criteria (the features which, when added, leads to the minimal criteria is chosen). We might observe in this figure that the deviance decreases monotonically as the model complexity increases. The same might not be said about the hit rate, although it has a growing pattern tendency, it has a decrease in the 6th and 7th iteration, when the features 4 and 2 are added.

If we consider the model obtained in the 5th iteration, it will consider the following features: 13, 12, 9, 8 and 3 (the meaning of those features may be seen in the table II). This 5th order model has achieved a 87% hit rate and is described by the following equation:

$$f(z) = \frac{1}{1 + e^{-z}}, \text{ where} \quad (26)$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5, \quad (27)$$

and the value of the parameters obtained are listed in the table III bellow

parameter	value
β_0	2.224
β_1	-0.542
β_2	-1.146
β_3	-1.146
β_4	0.0218
β_5	-0.477

TABLE III: Model's parameters value for the heart disease problem.

Doing an exhaustive search, creating all possible logistic models with all possible combination of features (a total of $\sum_{k=1}^n \binom{n}{k}$ possibilities, as $n = 13$ it gives us 8191 possibilities) we found the simplest model with the best hit-rate. This model comprises the following features: 1,2,3,4,5,8 and 13. For this model, the hit-rate was 0.889, the deviance 172.34 and the parameters are given in the table IV bellow.

parameter	value
β_0	0.25908
β_1	0.006829
β_2	-0.059335
β_3	-0.020367
β_4	0.032035
β_5	-0.713803
β_6	-0.858184
β_7	0.764581

TABLE IV: Best model's parameters for the heart disease problem.

The full model also gives the same hit-rate, but has 13 features instead of 7. The full model is the the model which leads to the minimum deviance, what show be expected by the definition of deviance.

C. Breast Cancer

This example application uses the data set of breast cancer from Wisconsin. The data are also available at the Machine Learning Repository (UCI, 2010). It consists of 30 features that were computed and describe characteristics of the cell nuclei present in the images from a fine needle aspirate (FNA) of a breast mass (see figure 5). The database has 569 samples which were used by Street et al. (1993) to extract the features listed in table V bellow.

The FNA provides a way to examine small amount of tumorous tissues by analyzing the characteristics of individual cells and important contextual features such as the size of cell clumps. This approach takes a long time from specialized physicians leading to a mixed success outcome. The process remains highly subjective, depending upon the skills and experience of the physician. The original work by Street et al. (1993) proposes a diagnosis method based on the analysis of the images from FNA, using image processing and machine learning techniques, what lead them to a 97% accurate method that is nowadays used in the University of Wisconsin Hospital to make breast cancer diagnosis, which is the second most deadly cancer in the U.S., since 1970 (Hoffman, 1993).

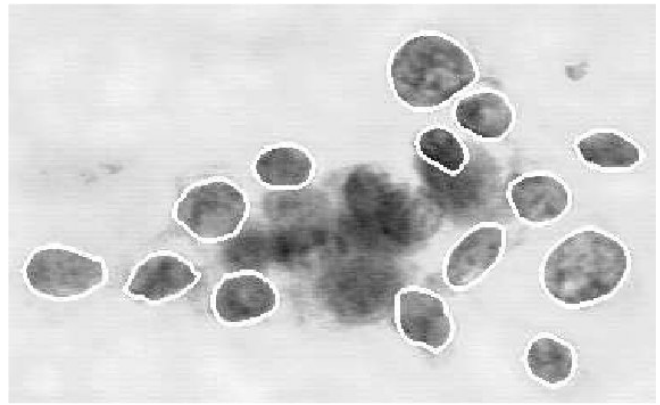


Fig. 5: This picture shows a magnified image of a malignant breast fine needle aspirate. The cell nuclei were outlined in white. The set of 30 features is extracted from sample images like this. This images is originally reproduced in Mangasarian et al. (1994).

id	feature
1, 2, 3	radius (mean, standard error, and largest)
4, 5, 6	texture (mean, standard error, and largest)
7, 8, 9	perimeter (mean, standard error, and largest)
10, 11, 12	area (mean, standard error, and largest)
13, 14, 15	smoothness (mean, standard error, and largest)
16, 17, 18	compactness (mean, standard error, and largest)
19, 20, 21	concavity (mean, standard error, and largest)
22, 23, 24	concave points (mean, standard error, and largest)
25, 26, 27	symmetry (mean, standard error, and largest)
28, 29, 30	fractal dimension (mean, standard error, and largest)

TABLE V: Data Set Features.

The figure 7b shows the results during the procedure of adding new features according to the minimal deviance criteria. Once more we observe that the deviance follows a monotonic decreasing pattern and the hit rate is between 0.88 and 0.95, although it might be considered a good result, it is always bellow the result shown by Street et al. (1993).

Considering the model achieved at the second iteration, only two features are used: 23 and 25 (standard error of the number of concave portions of the contour and symmetry mean value). The logistic model using those features results in a 93.8% hit rate and is expressed by the parameters:

parameter	value
β_0	31.893
β_1	-0.198
β_2	-78.493

TABLE VI: Model's parameters value for the breast cancer problem.

D. Chemotherapy-Sensitive Cancer

Selecting the best chemotherapy for an individual is of critical importance for this person treatment. A diagnostic test to guide selection of the optimal regimen for a particular patient is lacking. Pathologic complete response (pCR) after primary chemotherapy is associated with an excellent long-term prognosis, a complete eradication of all invasive cancer. The work proposed by Hess et al. (2006) aims at evaluating 'gene expression profiling as a potential tool to predict who may achieve pCR to sequential anthracycline paclitaxel preoperative chemotherapy'. The data provided by Hess et al. (2006) is made of 82 samples acquired from gene expression profiling performed with oligonucleotide microarrays (Affymetrix U133A) on

fine-needle aspiration specimens. The result of each microarray is the expression level of a 22,283 probe sets. More details on the data collection is provided by the authors Hess et al. (2006).

The problem proposed by this problem is similar to both previous problems shown above, but now the data has a huge number of features, and there are only a few samples available, what makes a very scarce sampling in the input space. That makes the problem rather difficult to solve. The same approach is used to analyze and classify the data using a logistic model. The usage of a MDS pre-step is really useful because it reduces drastically the computing time and leads to similar or better performance when comparing to the the situation when the features themselves are used as input for the logistic regression. The usage of MDS shows that it is possible to use the principal components of the representation of the samples as input for the logistic model and achieve better performance and reducing the computational cost. The results are illustrated in figure 9.

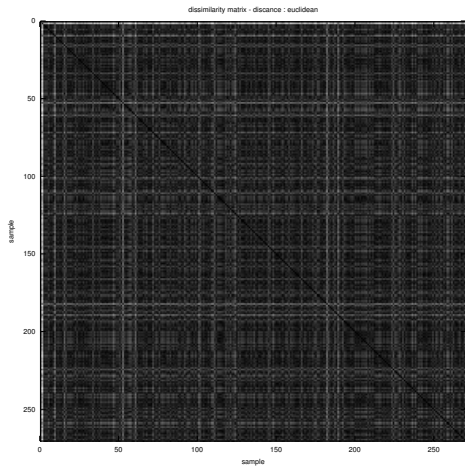
X. CONCLUSION

We have presented here the usage of a logistic regression in dichotomous classification tasks, illustrated by three examples (heart disease, breast cancer and chemotherapy sensitivity). For simpler problems, illustrated by the first two examples, it was shown that good results (hit-rate above 0.9) might be achieved with simple models. We also presented the usage of a multidimensional scaling as means of pre-processing the data, what leads to a small increase in performance. When dealing with complex problems involving a huge dimensionality and reduced observations, what leads to a scarce input space sampling, the pre-processing done by the multidimensional scaling also lead to a slight increase in hit-rate performance as well as a enormous decrease in computation algorithm cost, since it is cheap to calculate the distances between the small set of samples, and it further simplifies the creation of the logistic model, since now it involves a number of features (principal components) that is not greater than the number of samples (for the PCR problem presented, instead of dealing with the original 22,283 features, we would have to deal only with the 81 principal components resulted from the MDS, where 43 PCs would suffice to represent 90% of the data variance).

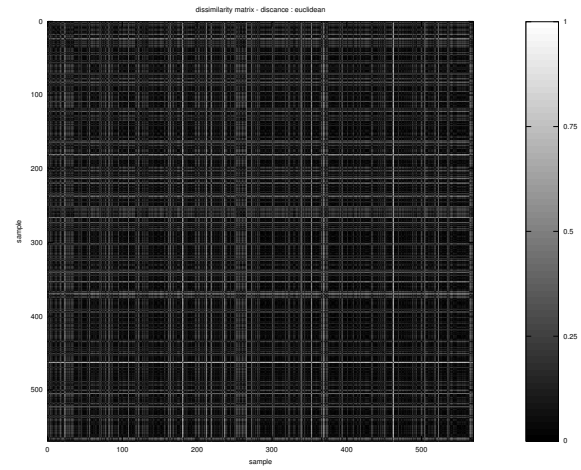
REFERENCES

- Cramer, J. (2003). The origins and development of the logit model. University of Amsterdam and Tinbergen Institute.
- Detrano, R. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64:304–310.
- Gennari, J. H., Langley, P., and Fisher, D. H. (1989). Models of incremental concept formation. *Artif. Intell*, 40:11.
- Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., Rouzier, R., Sneige, N., Ross, J. S., Vidaurre, T., Gomez, H. L., Hortobagyi, G. N., and Pusztai, L. (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, 24(26):4236–4244.
- Hoffman, M. S. (1993). *The World Almanac and Book of Facts 1993*. Pharos Books.
- Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage.
- Mangasarian, O. L., Street, W. N., and Wolberg, W. H. (1994). Breast cancer diagnosis and prognosis via linear programming. In *AAAI Spring Symposium on Artificial Intelligence in Medicine*, Stanford, CA.
- Street, W. N., Wolberg, W. H., and Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE*

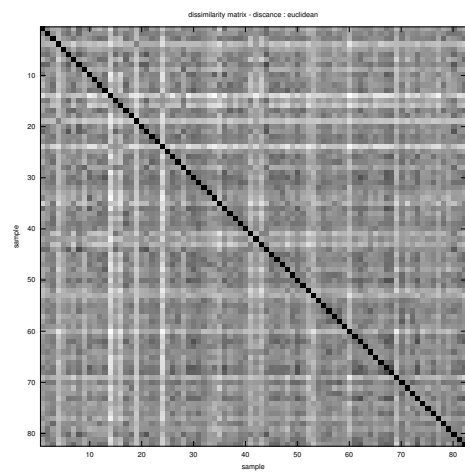
- 1993 International Symposium on Electronic Imaging: Science and Technology*, volume 1905, pages 861–870, San Jose, CA.
- Torgerson, W. (1952). Multidimensional scaling: 1. theory and method. *Psychometrika*, 17:401–419.
- UCI (2010). The uci machine learning repository. <http://archive.ics.uci.edu/ml/>.



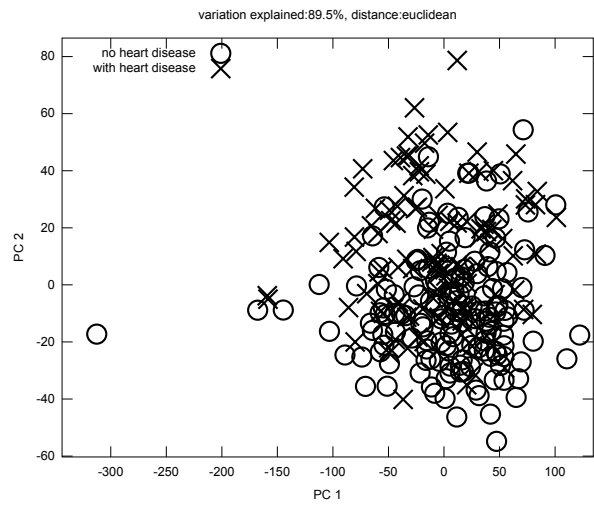
(a) Dissimilarity matrix for the heart disease database.



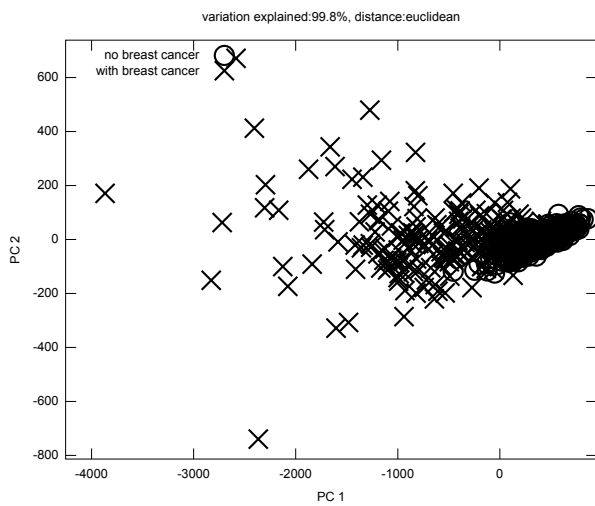
(b) Dissimilarity matrix for the breast cancer database.



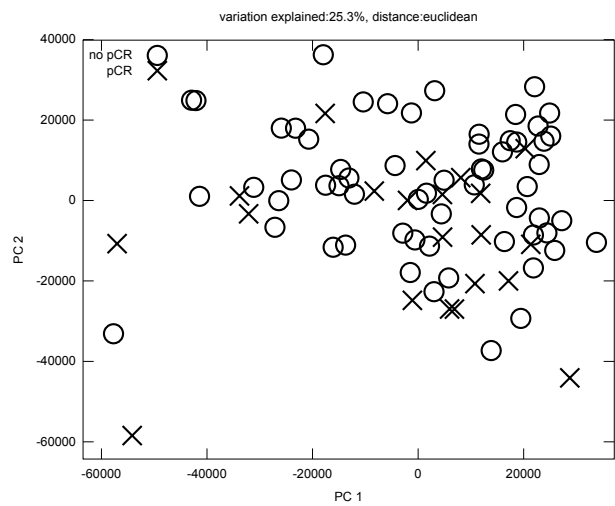
(c) Dissimilarity matrix for the PCR database.



(d) MDS result for the heart problem data. Plot of the data represented in the 2D space with the 2 principal components.

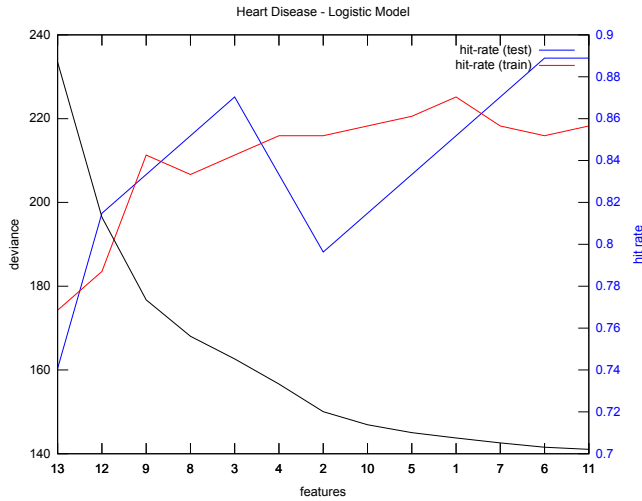


(e) MDS result for the breast cancer data. Plot of the data represented in the 2D space with the 2 principal components.

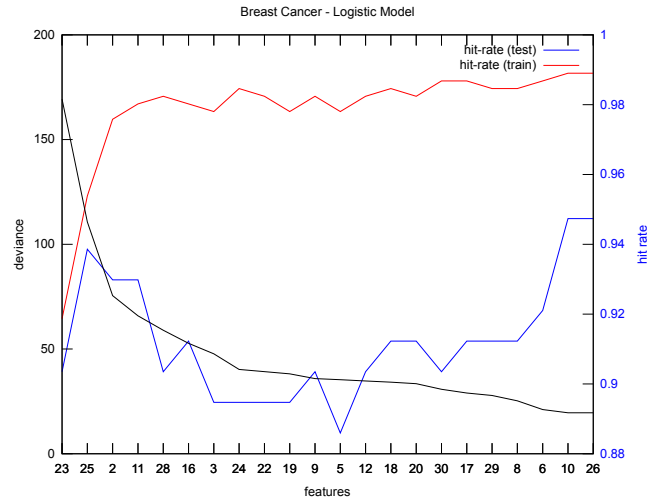


(f) MDS result for the PCR data. Plot of the data represented in the 2D space with the 2 principal components.

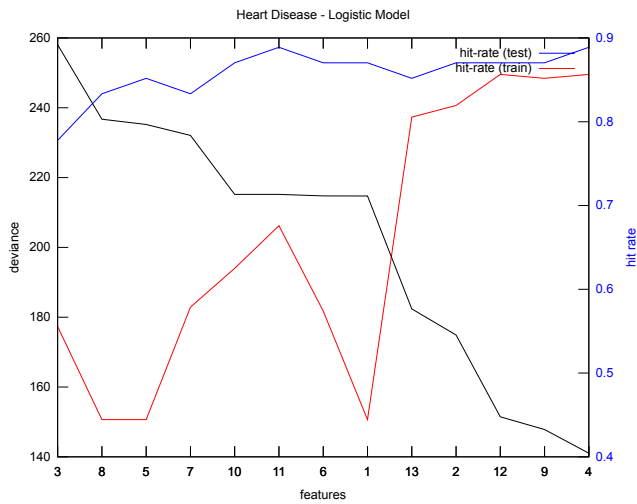
Fig. 6: The figures above show the dissimilarity matrix and the multidimensional scaling result for the data from the heart disease, breast cancer and PCR databases.



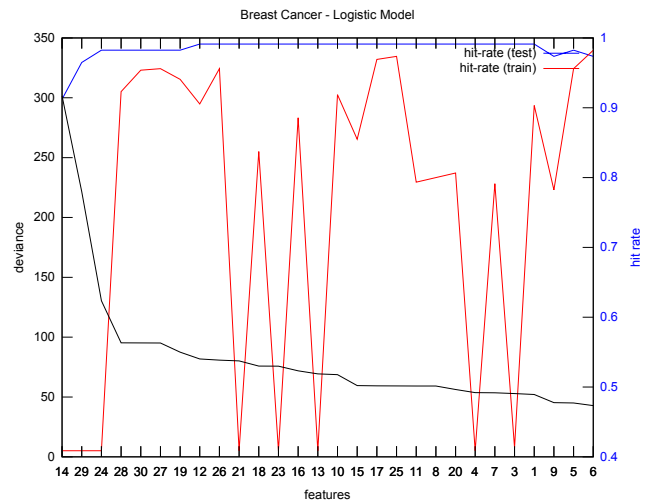
(a) Results for the heart disease problem when the features are added just according to a minimum deviance criteria.



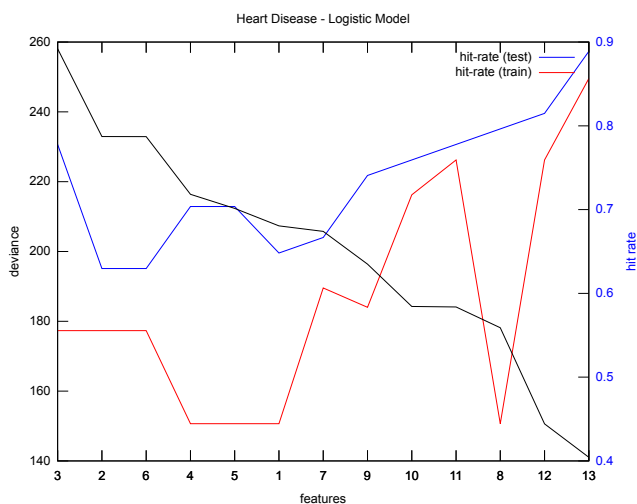
(b) Results for the breast cancer problem when the features are added just according to a minimum deviance criteria.



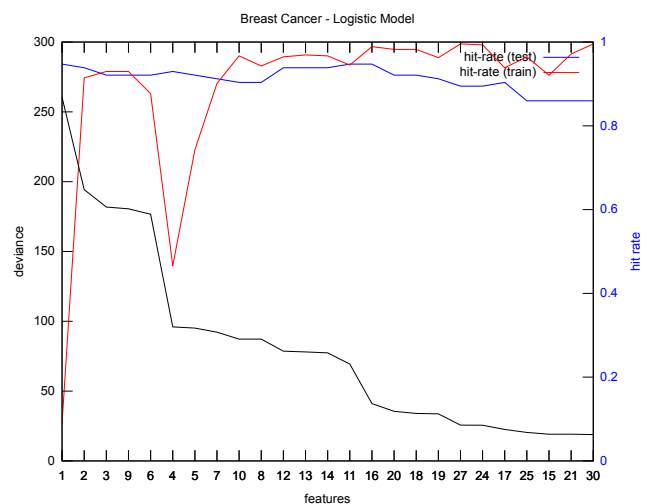
(c) Results for the heart disease problem when the features are added just according to a maximal hit-rate criteria.



(d) Results for the breast cancer problem when the features are added just according to a maximal hit-rate criteria.



(e) Results for the heart disease problem when the features are added just according to a maximal hit-rate criteria, constrained by the minimum deviance criteria.



(f) Results for the breast cancer problem when the features are added just according to a maximal hit-rate criteria, constrained by the minimum deviance criteria.

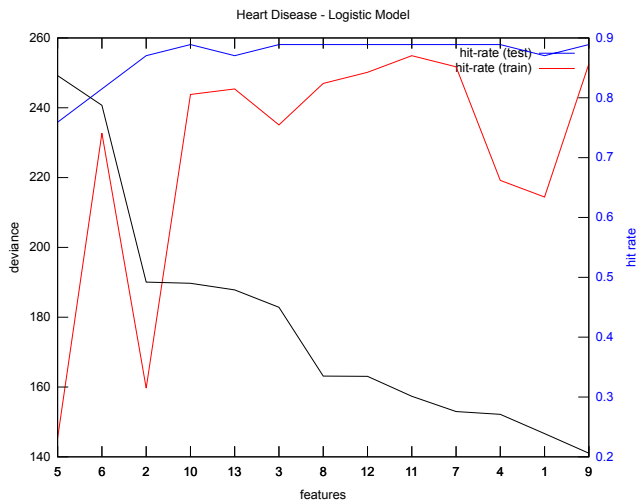
Fig. 7: The figures above show the deviance and hit-rate performance (shown in the y-axis) of a logistic regression. It starts with a simple model, a one feature model, and a new model is created at each iteration by adding a new feature (shown by the x-axis) to it. The results here displayed are for the heart disease problem and the breast cancer problem.



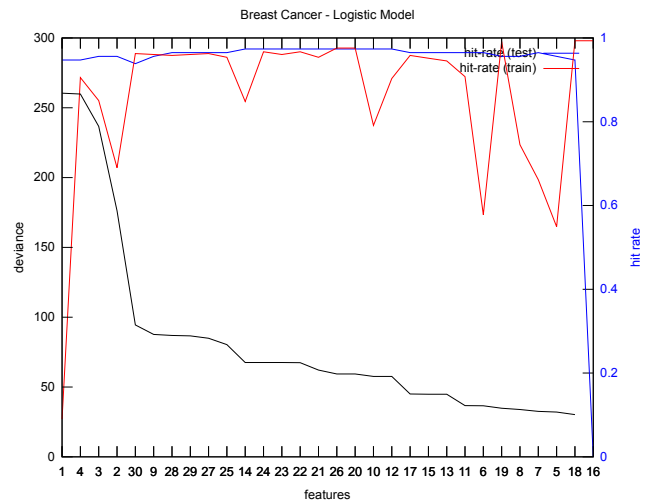
(a) Results for the heart disease problem when the principal components are added just according to a minimum deviance criteria.



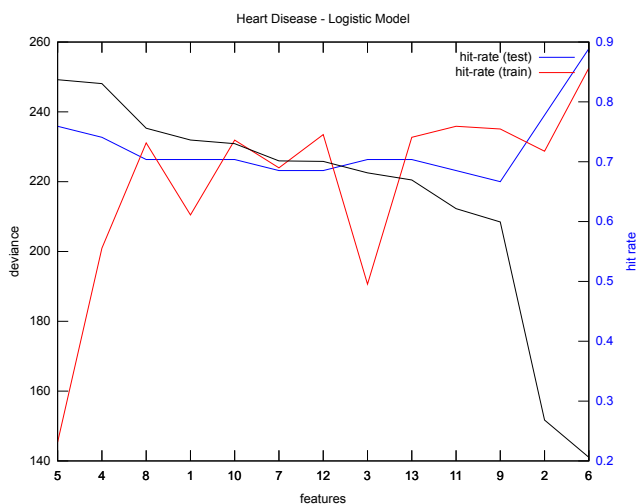
(b) Results for the breast cancer problem when the principal components are added just according to a minimum deviance criteria.



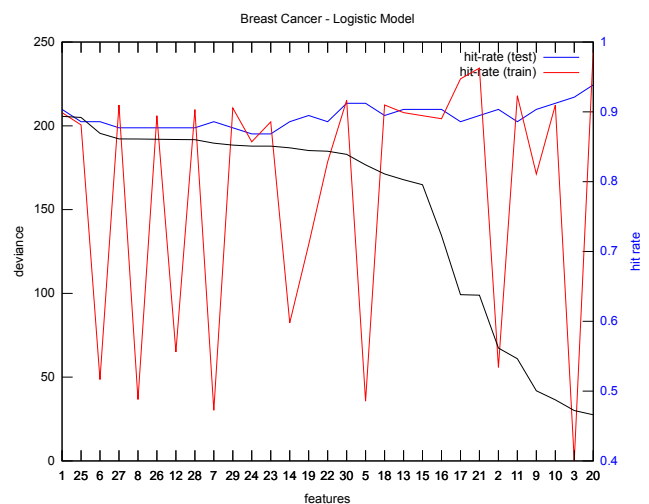
(c) Results for the heart disease problem when the principal components are added just according to a maximal hit-rate criteria.



(d) Results for the breast cancer problem when the principal components are added just according to a maximal hit-rate criteria.



(e) Results for the heart disease problem when the principal components are added just according to a maximal hit-rate criteria, constrained by the minimum deviance criteria.



(f) Results for the breast cancer problem when the principal components are added just according to a maximal hit-rate criteria, constrained by the minimum deviance criteria.

Fig. 8: The figures above show the deviance and hit-rate performance (shown in the y-axis) of a logistic regression. It starts with a simple model, a one feature (principal component, result of the MDS) model, and a new model is created at each iteration (shown by the x-axis) by adding a new feature (the other principal components from the MDS procedure) to it. The results here displayed are for the heart disease problem and the breast cancer problem when considering the MDS result of the dataset.

