Breast Cancer Detection

Executive summary

Using the Wisconsin Breast Cancer dataset, a Bayesian logistic regression model is constructed to predict whether or not a tissue sample is cancerous. After cleaning the data it was divided into training and test sets. Models of decreasing DIC were constructed using the training set and the final model, with smallest DIC, was tested on the test set where it achieved and accuracy of 95%.

Introduction

While breast X-rays can distinguish between a cyst and a tumor, they cannot distinguish between benign and malignant growths. For this purpose doctors often perform a major surgical procedure known as a biopsy. An alternative is a minor surgical procedure known as a fine-needle aspiration (FNA). In the latter case, the small sample size makes accurate diagnosis difficult. The purpose of this study is to determine whether it is possible to predict which lumps are malignant and which are benign based upon information obtainable from the smaller FNA sample, thus sparing women with benign tumors the stress of a full biopsy.

Data

The data used in this study comes from the Wisconsin Breast Cancer Database created by Dr. W. H. Wolberg. It consists of case studies from 699 patients collected over a three year period. From each patient an FNA sample of a suspicious lump was taken and nine characteristics visible in the digitalized images of each sample were measured. Dr. Wolberg then performed a full biopsy on each patient and classified the lumps as either benign or malignant. In all, the database contains 458 benign cases and 241 malignant cases.

Of the 699 cases, sixteen are lacking in one of the nine attributes. Fourteen of the incomplete cases are classified as benign and two are classified as malignant. In this report, the missing cases are dropped from the database leaving 683 cases. A change log in the database notes that a year after the data was taken, Dr. Wolberg changed two cases in which attribute values had been incorrectly entered.

The attributes take on values in the range [1,10], with larger values being indicative of cancerous conditions. The attribute distributions are shown in Fig. 1.

Fig. 2 figure shows the correlations between the attributes and the response variable, *cancer*. In agreement with the skewness of the distributions all the attributes are positively correlated with the response.

Note that *p2* and *p3* are strongly correlated, suggesting that one or the other could be dropped.

In this form the data is not well suited for modeling, so before proceeding the attributes are centered and scaled to lie in the range [-1,1].

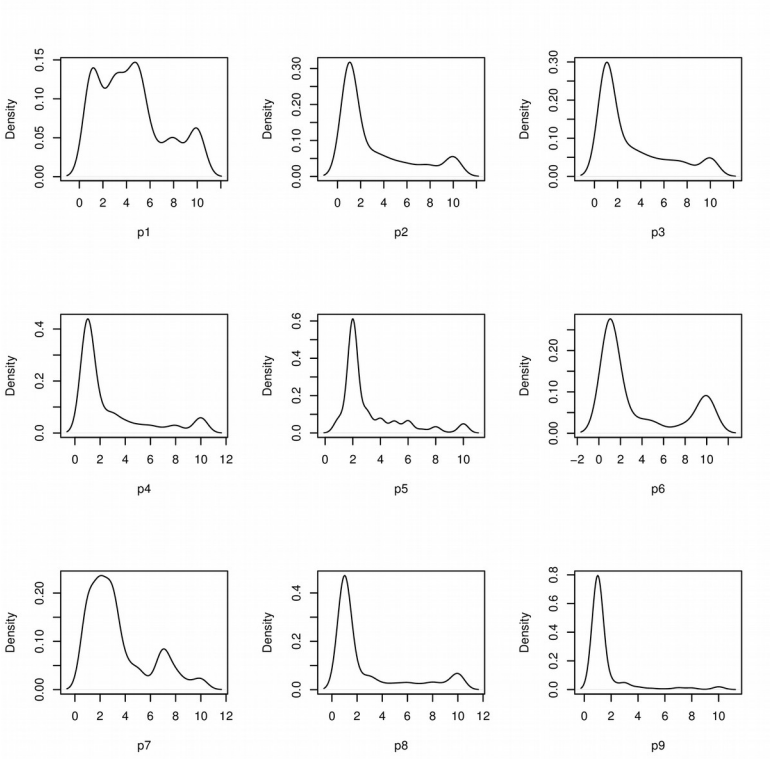

Fig. 1

Model

As the goal is two-level classification, logistic regression is the appropriate approach. Before starting, the data is randomly divided into a training set and a test set, whereby the ratio of positive to negative results in each split is the same as that found in the complete data set. The model will be built using the training set and its accuracy measured on the test set.

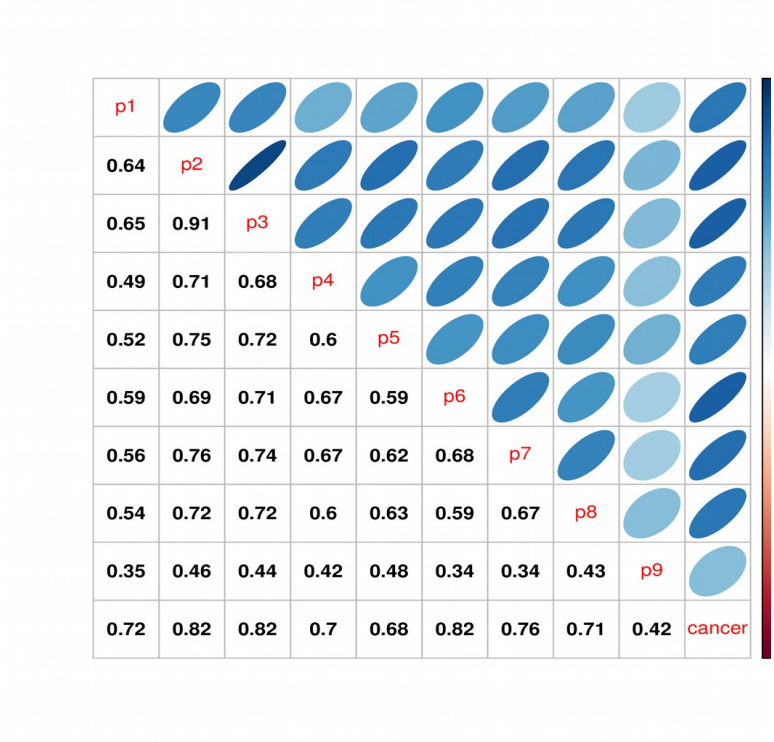

Fig. 2

A simple Bayesian model for logistic regression can be written in Rjags as:

```
model_I {
#Model
    for (i in 1:length(malig)) {
        malig[i] ~ dbern( phi[i] )
        logit(phi[i]) <- b0 + b[1]*p1[i] + b[2]*p2[i] + b[3]*p3[i] +
                              b[4]*p4[i] + b[5]*p5[i] + b[6]*p6[i] +
                              b[7]*p7[i] + b[8]*p8[i] + b[9]*p9[i]
    }
#Priors
    b0 ~ ddexp( 0.0, sqrt(2.0) )          # variance=1
    for ( j in 1:9 ) {
        b[j] ~ ddexp( 0.0, sqrt(2.0) )    # variance=1
    }
}
```

Model_I uses a double exponential prior with mean 0 and deviation 1 as we expect the regression coefficients to lie near zero. After initializing three chains, the model is run for 1,000 iterations beyond the point where autocorrelation graphs indicate all three chains have reached their stationary states. The chains are then run for a further 10,000 iterations to gather statistics. The resulting Gelman-Rubin's statistic is 1.0 for all predictors. The DIC for model_I is found to be 108.

Next, a hierarchical model is constructed in which the parameters for the priors are themselves drawn from a double exponential distribution.

```
model_II {
#Model
    for (i in 1:length(malig)) {
        malig[i] ~ dbern( phi[i] )
        logit(phi[i]) <- b0 + b[1]*p1[i] + b[2]*p2[i] + b[3]*p3[i] +
                              b[4]*p4[i] + b[5]*p5[i] + b[6]*p6[i] +
                              b[7]*p7[i] + b[8]*p8[i] + b[9]*p9[i]
    }
#Priors Level 1
    b0 ~ ddexp( mu, b1 )
    for ( j in 1:9 ) {
        b[j] ~ ddexp( mu, b1 )
    }
#Priors Level 2
    mu ~ ddexp( 0.0, sqrt(2.0) )      # variance=1
    b1 ~ dexp( 1.0 )                  # b1 > 0, variance=1
}
```

The same runs are made once again for the second model. The DIC for model_II is 106, which is smaller than the DIC for model_I, indicating it is the better model. However, a closer examination of the simulation results reveals that the posterior means of the coefficients for p2 and p5 are consistent with zero, indicating that these predictors can be dropped.

A third model, model_III, was then constructed with same hierarchy as model_II, but omitting p2 and p5. Following the same procedure as in the previous two runs, resulted in a DIC of 103. The posterior means of all the coefficients in model_III are positive and at least two standard deviations from zero; hence, further improvements by dropping predictors is unlikely.

Results

The main result is shown in Fig. 3, which depicts the accuracy of model_III as a function of the threshold, i.e., the value above which the posterior probability of cancer is considered indicative of a positive response. For production purposes one would choose the threshold that maximizes the accuracy on the training data which in this case is *T=0.2*.

In producing Fig. 3, the posterior means of the regression coefficients were used. Using the posterior estimates of the regression coefficients generated from each of the 30,000 runs to find the accuracy on the test set at *T=0.2*, yields a range of estimates, 95% of which lie in the interval (0.91, 0.96). Hence, we can be 95% confident that model_III's actual accuracy is within this interval.

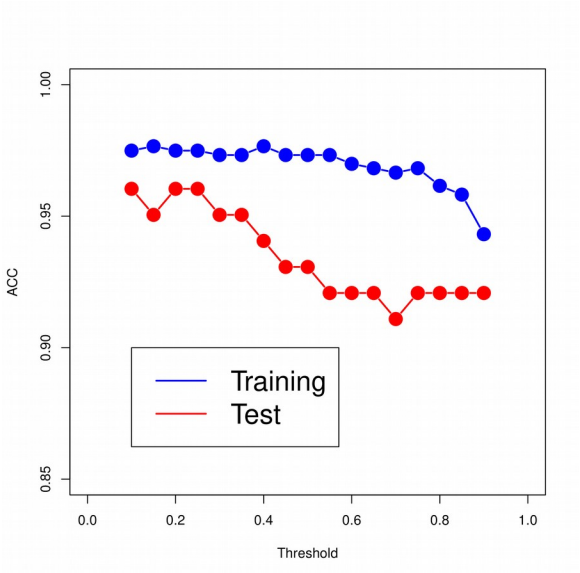

Fig. 3

Conclusions

A Bayesian regression model shows good promise for developing an expert system to determine whether a given FNA sample is cancerous. In this study, only one split of the data into test and training sets was used. To gain more confidence in the model's accuracy, multiple splits using a cross-validation scheme should be examined, an undertaking that was beyond the scope and means of this project.

Although prediction accuracy was used as a proxy to measure the model's performance, this may not be the best metric. It may be more desirable to optimize on the false-positive rate to prevent unnecessary surgical procedures, or to optimize on the false-negative rate to minimize the risk of cancer going undetected. However, such questions can only be answered in consultation with medical specialists.