# Application of Machine Learning Algorithms on Breast Cancer Dataset

Mochen Li[1], and Raji Sundararajan[2]

School of Engineering Technology, Purdue University, West Lafayette

Phone: (1) 765-430-5406

e-mail: li1049@purdue.edu, rsundara@purdue.edu

***Abstract-*** **Nowadays, with the continuous development of medical equipment, the size and features of medical database are also increasing rapidly. However, most current cancer detection still relies on doctors' observation and test with cell tissue samples which seems to have nothing to do with the above medical database. Therefore, here is a research focus for this: is there a pattern among all databases or how to find a pattern among these databases? Accordingly, data analysis solution or machine learning (ML) methods were applied more and more common among researchers and doctors. From Breast Cancer Surveillance Consortium (BCSC) dataset, 2,392,998 screening mammograms records are applied in this research. This dataset includes 14 independent variables and 2 dependent variables which can be called labeled data. Through developing the prediction model with generalized linear regression, naïve Bayes classifier, and support vector machine (SVM), a preliminary application and comparison between different machine learning models are presented. Among all three results, SVM shows the best accuracy for prediction but the longest computing time. Logistic regression has the bad performance with a large feature space but very high efficient. Here a simple machine learning algorithm selection process can be generated in these three methods:**

- **The first choice should be logistic regression. If its effect is not good then its result can be used as a reference.**
- **If the data size is small and aimed for multi-classification, naïve Bayes is more efficient and easier to implement.**
- **If the number of features and observations are particularly large, SVM will be a good option when resources and time are sufficient.**

## I. INTRODUCTION

Breast cancer is the top cancer among women around the world and also one of the leading causes of cancer death in women. In 2014, 236,968 women were diagnosed with breast cancer and 41,211 women died from it in the U.S [1] [2]. These data showed breast cancer accounts for the first and second place in female cancer incidence and mortality rate respectively. Meanwhile, 1 in 8 U.S. women have risk of developing breast cancer during her lifetime. From 1999 to 2014, it showed a continuous increase among the whole country and diagnosed population was raised with 49,007 new cases [2]. However, the incidence rate was going into a relative stable situation after decreasing for six consecutive years from 1999 to 2005. Studies have shown that old age is a risk factor for breast cancer, after developing an age-period-cohort analysis of 1999-2014 breast cancer dataset there was an interesting finding that women who were born between 1949 and 1954 usually got the highest incidence rate among all age-adjusted groups.

Currently, more and more data sets are collected for cancer diagnose and detection. And the utilization of machine learning in cancer research may be considered as an effective approach. Machine learning is a subfield of artificial intelligence (AI) that employs many algorithms allow computer learn from training data and develop a fitted pattern for prediction. This research presents a comparison among three machine learning algorithms applied in breast cancer dataset: support vector machine (SVM), naive Bayes classifier and generalized linear regression by measuring their accuracy, sensitivity and specificity values. [3]

## II. MACHINE LEARNING ALGORITM

Machine learning usually is defined as a subfield of artificial intelligence. Its aim can be generalized as to develop some algorithm that can learn from data and make “machine” predict the model of pattern by itself. In fact, there are too many machine learning algorithms can be used. In general, according to whether training data is labeled or unlabeled machine learning is usually categorized into two tasks: supervised learning and unsupervised learning. Besides, these two methods of machine learning have many algorithms to apply such as linear regression, decision tree, support vector machine etc. Obviously, all algorithms discussed in this research will be supervised learning. [4][5][6]

Generally speaking, every machine learning method has similar steps to solve problem.

1. Firstly, defining the objective problem. Through this step, it can quickly decide what kind of data research need or what kind of algorithm is more suitable.
2. Pre-processing of data. Before importing data into machine learning analysis, it usually requests formatting, cleaning and sometimes sampling for the data.
3. Algorithm selection. For supervised learning task, based on type of predictor variable task can be solved from two types: regression problem and classification problem. In this research, it is defined as a classification problem as the predictor can be described as have cancer (1) or no cancer (0). Meanwhile, classification problem also has many algorithms: support vector machine (SVM), naïve Bayes, logistic regression, decision tree, etc.
4. Finally, after applying the selected algorithm, it is necessary to assess and optimize the prediction model.

In this research, three basic algorithms will be discussed in the following sections: SVM, naïve Bayes classifier and generalized linear regression.

### *A. Support vector machine*

Generally speaking, SVM is a binary classifier which can be applied in both separable case and non-separable case. The SVM algorithm is to find a hyperplane that creates the biggest “margin” between two classes of training data ($y_i \in \{-1,1\}$). Figure 1 shows an example of SVM classifier in 3 dimensions.

For non-linear-separable case which is sometimes called non-linear classifier, it can be applied with Kernels SVM. Through Kernel functions, it maps the input vector from low dimensional space to the high dimensional space. Then it builds the hyperplane in high dimensional space to separate data. Here these feature mappings can be described as function φ. From linear separable case, it is known

$$\beta = \sum_{i=1}^{N} \alpha_i y_i x_i \tag{1}$$

then this can rewrite as

$$\mathrm{f(x)} = (\sum_{i=1}^{N} \alpha_i y_i x_i)^T x + \beta_0 = \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x\rangle + \beta_0 \tag{2}$$

When this function is mapped to high dimensional space, it can get:

$$\mathrm{f(x)} = \sum_{i=1}^{n} \alpha_i y_i \langle \varphi(x_i), \varphi(x)\rangle + \beta_0 \tag{3}$$

Therefore, kernel function can be defined like this:

$$\mathrm{K(x,z)} = \langle \varphi(x), \varphi(z)\rangle = \varphi(x)^T \varphi(z) = (x^T z)^2 \tag{4}$$

so, it just needs to calculate square of inner product of x and z. [4]

Based on risk factors defined by Centers for Disease Control and Prevention (CDC), 8 features were chosen from all 14 features for the following analysis [2]. 8 features and the output variable are defined in table 1 [7].

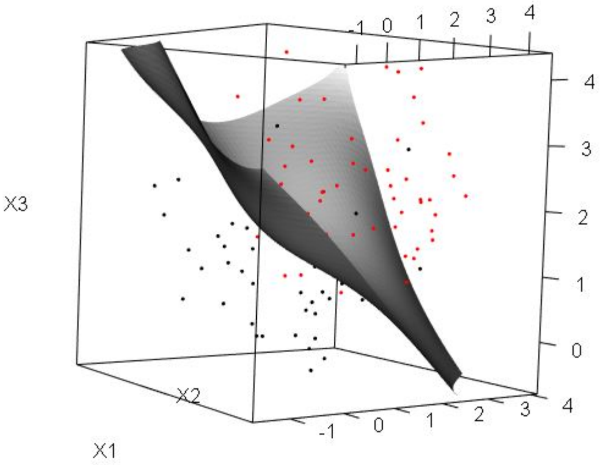
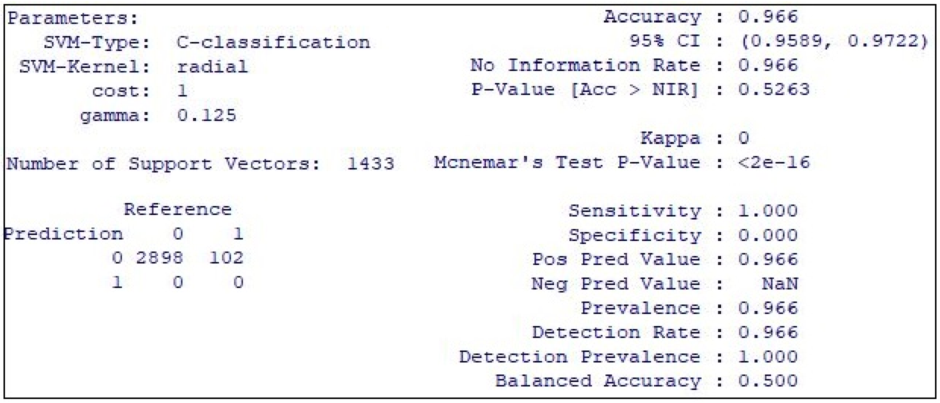


**Fig. 1 SVM plot in 3 dimensions**

TABLE 1: VARIABLES OF BREAST CANCER FROM BCSC

| | Variable | Coding |
|---|---|---|
| $x_1$ | Menopaus | 0=premenopausal, 1=postmenopausal |
| $x_2$ | Agegrp | Age group: 1=35-39; 2=40-44; 3=45-49; 4=50-54; 5=55-59; 6=60-64;<br>7=65-69; 8=70-74; 9=75-79; 10=80-84 |
| $x_3$ | Density | 1 = Almost entirely fat; 2 = Scattered fibroglandular densities; 3 = Heterogeneously dense; 4 = Extremely dense |
| $x_4$ | Race | 1= white; 2=Asian/Pacific Islander; 3=black; 4=Native American; 5=other/mixed |
| $x_5$ | Bmi | Body mass index: 1=10-24.99; 2= 25-29.99; 3=30-34.99; 4=35 or more |
| $x_6$ | Agefirst | Age at first birth: 0=Age<30; 1=Age 30 or greater; 2=Nulliparous |
| $x_7$ | Nrelbc | Number of first degree relative with breast cancer: 0=zero; 1=one;<br>2=2 or more |
| $x_8$ | Hrt | Current hormone therapy: 0=no; 1=yes |
| y | invasive | Diagnosis of invasive breast cancer within one year of the index screening mammogram: 0=no; 1=yes |

In order to testing model, dataset usually is split into three parts: training data (60%), test data (20%), and validation data (20%). After applying breast cancer risk factor dataset into this SVM model, the final result is shown in Figure 2.



```
Parameters:                                              Accuracy : 0.966
   SVM-Type:  C-classification                             95% CI : (0.9589, 0.9722)
 SVM-Kernel:  radial                          No Information Rate : 0.966
       cost:  1                               P-Value [Acc > NIR] : 0.5263
      gamma:  0.125
                                                            Kappa : 0
Number of Support Vectors:  1433   Mcnemar's Test P-Value : <2e-16

          Reference                                   Sensitivity : 1.000
Prediction    0    1                                  Specificity : 0.000
         0 2898  102                               Pos Pred Value : 0.966
         1    0    0                               Neg Pred Value :   NaN
                                                       Prevalence : 0.966
                                                   Detection Rate : 0.966
                                             Detection Prevalence : 1.000
                                                Balanced Accuracy : 0.500
```

Fig. 2. SVM classifier result

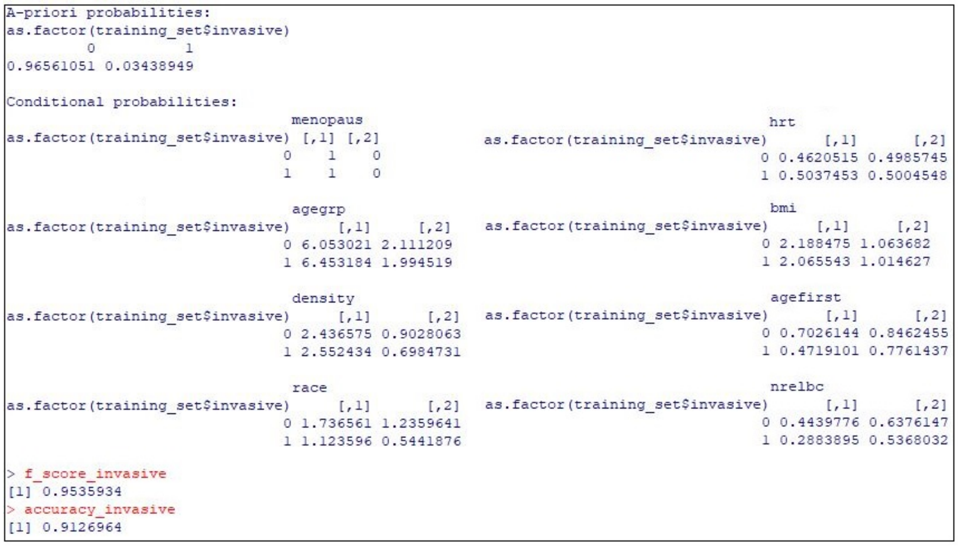
### *B. Naïve Bayes classifier*

Naïve Bayes method is a classification method based on Bayesian theorem and feature condition independent hypothesis. For a given training dataset, the joint probability distribution of the input/output is first learned based on independent assumptions of feature conditions; then according to this model it uses Bayes theorem to find the output $y$ with the largest posterior probability for a given input $x$. Based on this assumption, the Bayesian formula can be rewritten as:

$$\mathrm{p(c|x)} = \frac{p(c)p(\mathcal{X}|\mathcal{C})}{p(x)} = \frac{p(c)}{p(x)}\prod_{i=1}^{d} p(x_i|c) \qquad (5)$$

here $d$ is the number of features, $x_i$ is the value of $i_{th}$ feature.

With the same dataset from BCSC, naïve Bayes classifier result is shown in Figure 3.



```
A-priori probabilities:
as.factor(training_set$invasive)
         0          1
0.96561051 0.03438949

Conditional probabilities:
                                menopaus                                          hrt
as.factor(training_set$invasive) [,1] [,2]        as.factor(training_set$invasive)         [,1]      [,2]
                               0    1    0                                       0 0.4620515 0.4985745
                               1    1    0                                       1 0.5037453 0.5004548

                                agegrp                                            bmi
as.factor(training_set$invasive)     [,1]     [,2]  as.factor(training_set$invasive)       [,1]     [,2]
                               0 6.053021 2.111209                                 0 2.188475 1.063682
                               1 6.453184 1.994519                                 1 2.065543 1.014627

                                density                                           agefirst
as.factor(training_set$invasive)     [,1]      [,2]  as.factor(training_set$invasive)        [,1]      [,2]
                               0 2.436575 0.9028063                                 0 0.7026144 0.8462455
                               1 2.552434 0.6984731                                 1 0.4719101 0.7761437

                                race                                              nrelbc
as.factor(training_set$invasive)     [,1]      [,2]  as.factor(training_set$invasive)        [,1]      [,2]
                               0 1.736561 1.2359641                                 0 0.4439776 0.6376147
                               1 1.123596 0.5441876                                 1 0.2883895 0.5368032

> f_score_invasive
[1] 0.9535934
> accuracy_invasive
[1] 0.9126964
```

**Fig. 3. Naive Bayes classifier result**

### *C. Generalized linear regression (Logistic regression)*

There are actually many similarities between logistic regression and multiple linear regression. The biggest difference is that their dependent variables are different types. In fact, both of them can be attributed to the same family which is named generalized linear regression (GLR). [5] [6] [8]

Models of GLR are basically the same, except the difference of dependent variables among models:

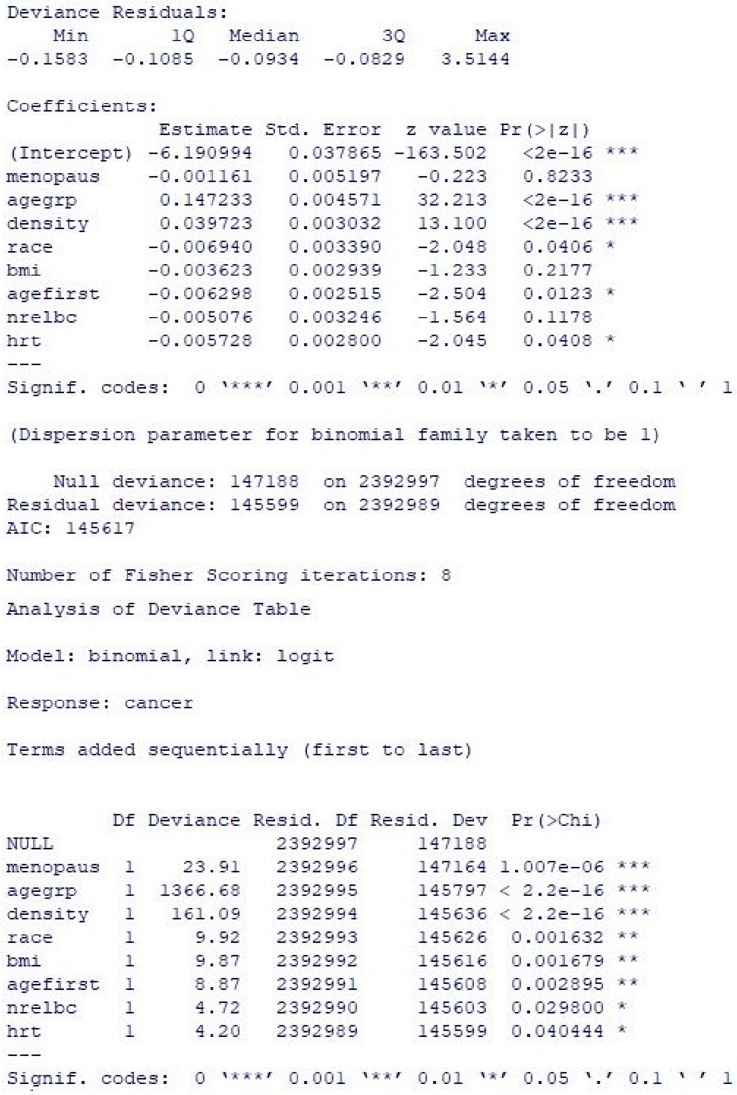
- If dependent variable is continuous, this is a multiple linear regression.
- If dependent variable is a binomial distribution, this is a logistic regression.
- If dependent variable is a Poisson distribution, this is a Poisson regression.

- If dependent variable is a negative binomial distribution, it is a negative binomial regression.

Logistic regression formula can be written as:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \tag{6}$$

Because the dependent variable here is a binary value, the prediction model uses logistic regression. The final result of logistic regression analysis is shown in Figure 4.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1583  -0.1085  -0.0934  -0.0829   3.5144

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -6.190994   0.037865 -163.502   <2e-16 ***
menopaus    -0.001161   0.005197   -0.223   0.8233
agegrp       0.147233   0.004571   32.213   <2e-16 ***
density      0.039723   0.003032   13.100   <2e-16 ***
race        -0.006940   0.003390   -2.048   0.0406 *
bmi         -0.003623   0.002939   -1.233   0.2177
agefirst    -0.006298   0.002515   -2.504   0.0123 *
nrelbc      -0.005076   0.003246   -1.564   0.1178
hrt         -0.005728   0.002800   -2.045   0.0408 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 147188  on 2392997  degrees of freedom
Residual deviance: 145599  on 2392989  degrees of freedom
AIC: 145617

Number of Fisher Scoring iterations: 8
Analysis of Deviance Table

Model: binomial, link: logit

Response: cancer

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                   2392997     147188
menopaus  1    23.91   2392996     147164 1.007e-06 ***
agegrp    1  1366.68   2392995     145797 < 2.2e-16 ***
density   1   161.09   2392994     145636 < 2.2e-16 ***
race      1     9.92   2392993     145626  0.001632 **
bmi       1     9.87   2392992     145616  0.001679 **
agefirst  1     8.87   2392991     145608  0.002895 **
nrelbc    1     4.72   2392990     145603  0.029800 *
hrt       1     4.20   2392989     145599  0.040444 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Fig. 4. Logistic regression result**

## III. RESULTS AND DISCUSSION

This research has explored 8 features for prediction model with three basic machine learning algorithms: SVM, naïve Bayes classifier and logistic regression. From three result sheets, SVM showed the best accuracy. However, each machine learning algorithm has their limitations. Because of the huge cost of computing in SVM algorithm, only 10,000 sampled data were applied in SVM prediction model. In spite of these quite small number, SVM algorithm cost much longer time than other two methods in learning data and resulted in 1433 support vectors. On the other hand, sometimes it is very hard to decide a good kernel function in applying kernel SVM. By comparison, naïve Bayes method has a slightly worse classification, but it is highly efficient and easy to implement. It seems that logistic regression has the lowest accuracy among three methods. However, logistic regression still has many applications in cancer research. In fact, logistic regression is very convenient and its most useful characteristic is that the output is not a discrete value but a probability which is associated with the observed sample. Based on this, combining different given thresholds, it can generate some best classification result to meet the problem.

## REFERENCES

[1] Breast cancer: prevention and control, WHO, http://www.who.int/cancer/detection/breastcancer/en/
[2] Breast Cancer, Centers for Disease and Prevention, https://www.cdc.gov/cancer/breast/
[3] Ahmad LG, Eshlaghy AT, Poorebrahimi a, & Razavi AR, Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence, Health & Medical Informatics, 2013, 4 (2)
[4] Tom M. Mitchell, Machine Learning. McGraw Hill, ISBN 0-07-042807-7
[5] Trevor Hastie, Robert Tibshirani, & Jerome Friedman, The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Springer, ISBN: 978-0-387-84857-0
[6] Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, ISBN: 978-1-4614-7137-0
[7] Risk Factors Dataset, Breast Cancer Surveillance Consortium, http://www.bcsc-research.org/data/
[8] Delen D, Walker G, Kadam A, Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, 2005, 34, pp113-127