Breast Cancer Prediction – Benign or Malignant (April 2019)

Martin Garcia, marting@smu.edu, Andrew L. Wilkins, awilkins@mail.smu.edu

***Abstract*—Data was collected from the UCI Machine Learning Respository for Breast Cancer. The data includes multiple explanatory variables of the cancer cell being observed and a binary response variable of malignant or benign.**

***Index Terms*—Cancer, Breast Cancer, malignant, benign, classification model, healthcare, supervised classification, discriminant analysis, principal component analysis, binary response**

I. Introduction

THE goal of the case study is to accurately predict whether a cancer cell is malignant or benign based on 30 different explanatory variables. Malignancy is a general term for a cell that divides uncontrollably and spreads. These rogue cells have various names that are dependent on the location they form. Various examples include Carcinoma for malignancy that starts in the skin and Sarcoma which begins in bones, cartilage, fat, or blood vessels. Breast cancer is usually categorized as a carcinoma. In contrast, benign cells don’t have the ability to spread to their neighbors. (See figure one)

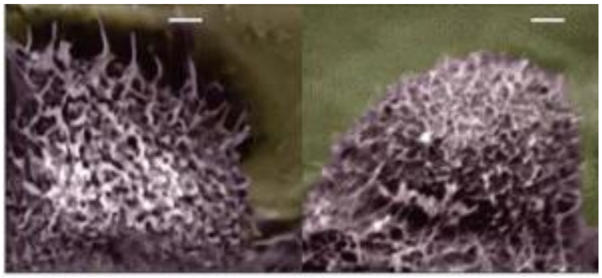


Figure 1 - Can you guess which cell is malignant? The images above were captured with an atomic force microscope capable of capturing roughly a nanometer in size.

II. Data Description

Our dataset was pulled from the UCI Machine Learning Repository. This data was collected from 1989 to 1991 at the University Of Wisconsin Hospital in Madison. The explanatory variables were measurements taken from images of cell nuclei of breast mass samples. This includes 569 observations paired with continuous variables that will help us predict a binary response variable, malignant or benign as 0 or 1 respectively. The data consist of physical measurements of the cell such as area, perimeter and many others (see figure 2). Observations are divided into 357 benign and 212 malignant cases.

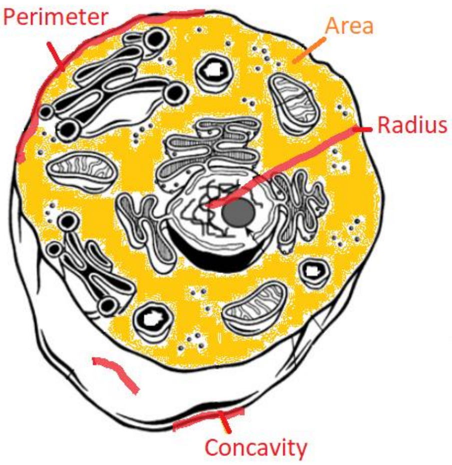


Figure 2 - Parameter examples used for our analysis

III. Exploratory data analysis

To gain a better understanding of the data we generated histogram and qq plots for every explanatory variable. To meet the assumptions of our testing methods we had to look at the distributions, variance, independence, outliers, and linearity. The above stated assumptions meet the requirements for principal component analysis (PCA) and logistic regression analysis.

A. Distribution

Data sets were not individually normally distributed and showed skewed results for several explanatory variables. This was addressed by taking the log of these variables. However, our sample size is large enough to be robust against skewness thanks to the central limit theorem.

The above criteria only addresses the normality assumption for PCA. Linear regression requires a univariate normal distribution for our response variable and multivariate normal distribution for our explanatory variables. When we plot each principle component against one another we see that the 95% confidence ellipses form circles around the data. (see figure 2)

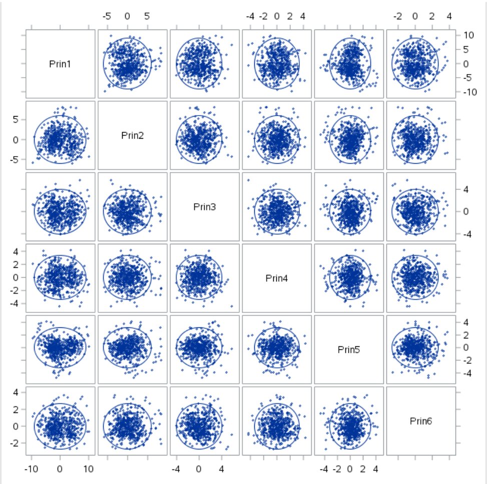


Figure 3 - Multivariate distribution of principle components

A model of our response variable in a univariate normal distribution would result in an ‘S’ shaped distribution. This was addressed by using a logit function or general linear model through our logistic regression. (see figure 4)

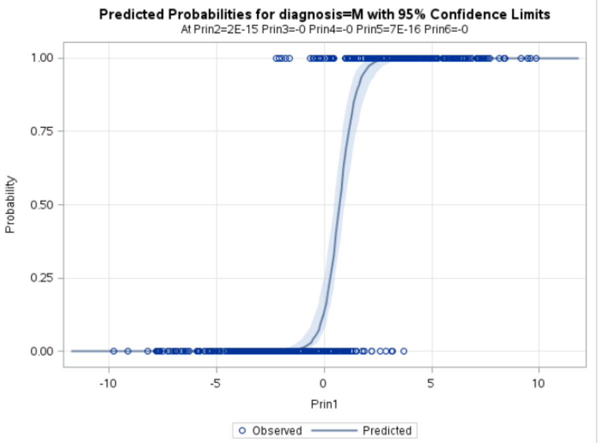


Figure 4 - Univariate distribution of response variable

B. Variance

Variances for the area mean and area worst variables were exponentially larger compared to the rest. After performing a log transformation and adjusting for scaling through PCA our variances no longer showed large absolute differences.

Logistic regression requires homogeneity across all variances of each response variable and all explanatory variables. Since we are performing a dimension reduction through principal component analysis, this assumption is satisfied through the log transformation of the explanatory variables that make up the principal components.

C. Independence

Independence is assumed since the patients selected in this study have no connection to one another other than being prior patients of Dr. Wolberg.

D. Linearity

Linearity for PCA – relationships between all observed variables should be linear.

After performing some exploratory data analysis and examining the scattermatrix of the logarithmic transformation of each explanatory variable, we find that the linearity assumption is in fact met.

Given the fact that this study is retrospective, we will proceed with our analysis using odds ratios.

IV. Statistical Analysis

A. Principal Component Analysis

By employing principle component analysis we were able to reduce the number of explanatory variables from 30 down to 6. Variances were substantially different from each other across the 30 explanatory variables. Therefore a standardized approach to PCA seemed appropriate over a covariance matrix. Standardization helped us since no one explanatory variable was more important than another. This is important because PCA will emphasize variables with high variances and we want to capture the variables that account for the overall large proportion of variance. To examine the variance, we start with summary statistics of all variables and check if any show large cumulative variation compared to the rest or stand out as outliers. Two outliers were present in the area mean and area worst variables and suggest a correlation matrix, confirming the necessity for standardization.

We tried 3 different approaches: using a log transformation on a correlation matrix, using a log transformation on the covariance matrix, and using a correlation matrix without the log transformation. Transforming our data not only normalized our variables, but it also helped us compare variance differences in absolute terms to allow for accurate comparison of variance for our explanatory variables. Based on this, we believe a doubling of our scaling through log transformation and scaling method provided by correlation matrix was our best approach.

The first six principle components account for 90% of the variance (see figure 4). During each iteration of the principle component analysis a combination of our variables that account for the most variance is consolidated as a new variable, or principle component.

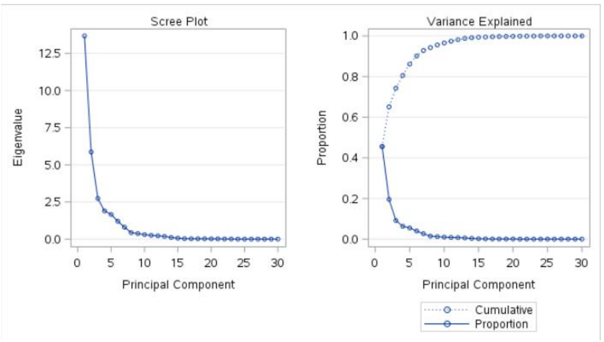


Figure 4 - Scree Plot suggest most variability is accounted for before the 6th principal component

The total variance is reflected in our eigenvalues and eigenvectors function as coefficients for our linear combinations of explanatory variables.

Translating the principal components

Translating loadings in relation to principle components in relation to a linear regression are better demonstrated using the principle components and their respective loadings. Loadings help us determine the level of impact for a particular component. Loadings ranged from 0.2 to 0.6 for each principle component.

As an example, we focus on only principle component 1 to explain the effects of our explanatory variables. No one explanatory variable accounts for all the variability. The average loading for principle component 1 was around 0.2. Variables that help explain the variance in probability of our model include compactness mean, concavity mean, concave points mean, and concave points worst. This suggests that principle component 1 primarily focuses on the concavity of a cancer cell (see figure 5).

Similarly, principle component 2 tells us a different story as this variable shows the most variance for a different set of explanatory variables such as fractal dimension mean and standard error. Loadings can hold positive and negative values which indicate whether these explanatory variables lower or raise the probabilities for this particular model.

If we interpret only principle component 1, this would mean positive loadings increase the probability of a cell being malignant. We would then look at those 4 variables previously discussed with high loadings as those mainly responsible for the variability in our response, or probability. Then, we would place our attention at compactness of the cell, mean of concave portions of our cell, mean number of concave points, and most average number of concave points.

Interpreting the variates introduced by principal component analysis requires us to analyze the loadings of each of the variables contained within. The 6 chosen variates account for over 90% of the variability. We believe this was a good benchmark to predict if a cancer cell is benign.

To illustrate this, Figure 5 contains the most signicant loadings per principle component with red bars signifying negative loadings and blue for positive. As described above, Prin1 focuses on the concavity of the cell. Prin2 is concerned with the fractal dimension. Prin3 emphasizes texture and smoothness of the cancer cell, while Prin4 places even more emphasis on texture. Prin5 takes contrasts the symmetry of the cell versus the concavity of it. Finally, Prin6 contrasts the symmetry against the smoothness. The collection of these 6 variates will then in turn predict whether a cancer cell is malignant or benign.

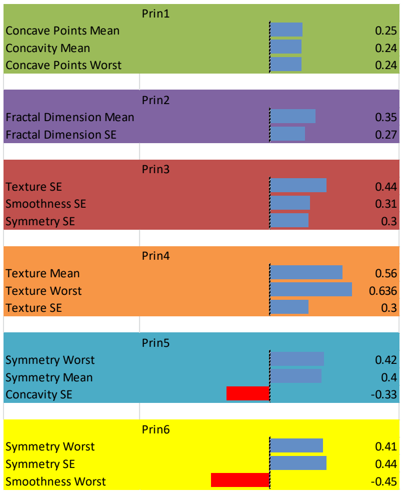
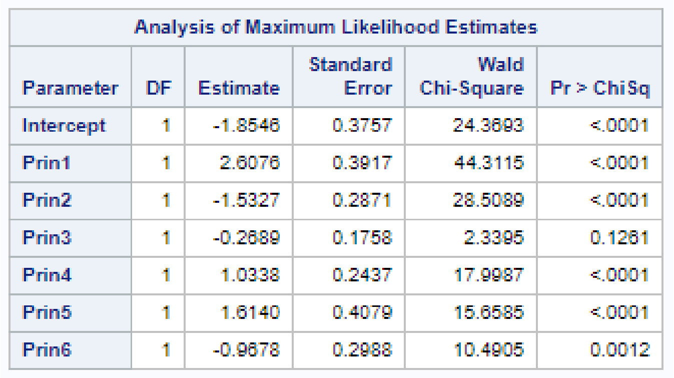


Figure 5 - PCA and top 3 loadings example. Blue bars indicate positive loadings while red indicates negative loadings.

B. *Logistic Regression*

The response for our case study is binary with malignant or benign as the only two options. To accommodate for multiple explanatory variables we use a correlation matrix to determine what attributes most contribute to benign cancer.

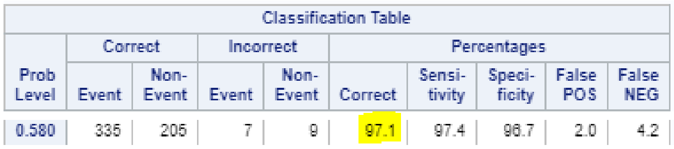
We use the method of maximum likelihood in place of ordinary least squares. Instead of determining a linear relationship of our data points based on a minimum of our residuals, we determine our coefficients for logistic regression based on those that contribute to the largest probability of benign cancer. Results are all are statistically significant with the exception of principle component 3 (see figure 6). We use the Wald Chi-Square as our test statistic. Our Hosmer and Lemeshow Goodness-of-Fit Test has a p-value = 0.9335. This indicates our model may not be a poor fit.



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8546	0.3757	24.3693	<.0001
Prin1	1	2.6076	0.3917	44.3115	<.0001
Prin2	1	-1.5327	0.2871	28.5089	<.0001
Prin3	1	-0.2689	0.1758	2.3395	0.1261
Prin4	1	1.0338	0.2437	17.9987	<.0001
Prin5	1	1.6140	0.4079	15.6585	<.0001
Prin6	1	-0.9678	0.2988	10.4905	0.0012

Figure 6 - Principle Component Analysis with resulting 6 variables

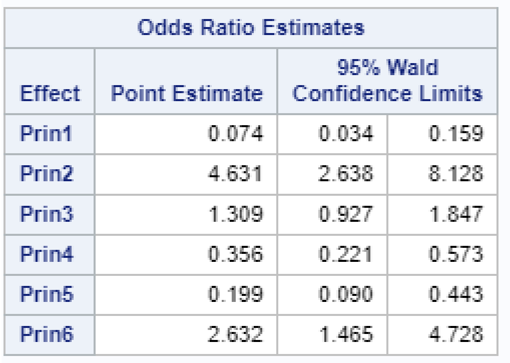
Our classification table (see figure 7) produced the highest percentage correctly classified at 97.1% at a probability level of 58% that the cell is benign. 97.1% translates as a measure of accuracy per the set probability level. If probability level is exceeded, the cancer is classified as benign. Otherwise, the classification of the cancer is malignant. Sensitivity at 97.4% translates into our model predicting benign cancer correctly when our cell is actually benign. Similarly, our model correctly predicts malignant cancer in 96.7% of all cases where the true classification is malignant. However, using sensitivity and specificity in this manner only applies to an observational study. Therefore, we must interpret our model in terms of odds ratios.



Classification Table									
	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.580	335	205	7	9	97.1	97.4	96.7	2.0	4.2

Figure 7 - Classification table at 58% probability level

Our odds ratio estimates (Figure 8) display the increase in the odds of benign cancer cells compared to malignant cancer cells for each principle component. Significant increases in odds are noticed in the 2nd principle component, indicating a one unit change in Prin2 increases the odds of benign cancer by 4.631 holding all other variates constant.



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Prin1	0.074	0.034	0.159
Prin2	4.631	2.638	8.128
Prin3	1.309	0.927	1.847
Prin4	0.356	0.221	0.573
Prin5	0.199	0.090	0.443
Prin6	2.632	1.465	4.728

Figure 8 - Odds ratio for benign cells

Recall, Prin2 emphasizes fractal dimension of the cancer cell. Therefore, a larger mean fractal dimension and standard error of the fractal dimension signifies the greatest increase in the odds of the cancer being classified as benign. The next largest increase in the odds of benign cancer classification comes from principal component 6. This variate tells us that the contrast between the worst symmetry and the worst smoothness a key components in determining the classification of the cancer.

We also test the probability that at least one of our coefficients of our model is not equal to zero using the likelihood ratio. A redundant test is also done with Wald's test which both show pr < 0.0001.

V. Conclusion

Reducing the number of explanatory variables through PCA and creating a linear regression model that uses a log transformation resulted in a better interpretation of the significance of the variables. We can calculate the weight of each principle component and the constituents that make them up for our explanatory variables by weighting them based on the proportion of variance they explain (see figure 10).

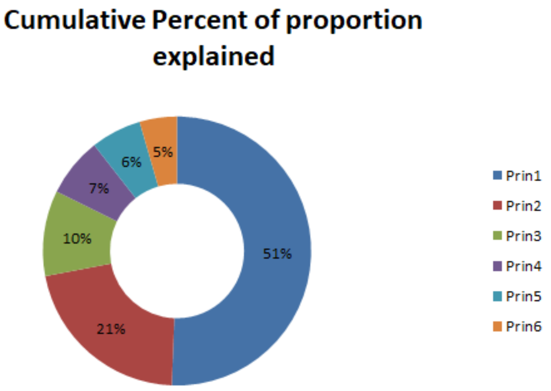


Figure 10 - Principle components broken per cumulative percent

Each individual variate can be broken down by the explanatory variables that comprise them (see figure 11). The sum of the weights of each explanatory variable at each iteration of the principle component helps best explain the model.

Figure 11 - each color represents a different explanatory variables weight for principle component 1

This translates in terms of odds due to our binary response which can be calculated in many ways holding specific principle components constant while measuring others.

VI. References

[1] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/activity

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] https://www.cancer.org/

[4] https://www.nature.com/articles/nnano.2009.77

VII. Appendix

SAS Code:

```
data Cancer;
infile '/home/marting0/Data/data.csv' firstobs=2
dlm=",";
input id diagnosis $  radius_mean  texture_mean
  perimeter_mean area_mean
      smoothness_mean  compactness_mean
  concavity_mean concave_points_mean
      symmetry_mean  fractal_dimension_mean
  radius_se  texture_se perimeter_se
```

```
      area_se  smoothness_se  compactness_se
  concavity_se concave_points_se
      symmetry_se  fractal_dimension_se radius_worst
  texture_worst
      perimeter_worst  area_worst smoothness_worst
  compactness_worst
      concavity_worst  concave_points_worst
  symmetry_worst fractal_dimension_worst;
run;
proc print data= cancer; run;

/* The data are not normally distributed.  Take the
log of each parameter and replot.; */

data Cancer2;
set Cancer;
radius_mean = log(radius_mean);
texture_mean = log(texture_mean);
perimeter_mean = log(perimeter_mean);
area_mean = log(area_mean);
smoothness_mean = log(smoothness_mean);
compactness_mean = log(compactness_mean);
concavity_mean = log(concavity_mean);
concave_points_mean = log(concave_points_mean);
symmetry_mean = log(symmetry_mean);
fractal_dimension_mean =
log(fractal_dimension_mean);
radius_se = log(radius_se);
texture_se = log(texture_se);
perimeter_se = log(perimeter_se);
area_se = log(area_se);
smoothness_se = log(smoothness_se);
compactness_se = log(compactness_se);
concavity_se = log(concavity_se);
concave_points_se = log(concave_points_se);
symmetry_se = log(symmetry_se);
fractal_dimension_se = log(fractal_dimension_se);
radius_worst = log(radius_worst);
texture_worst = log(texture_worst);
perimeter_worst = log(perimeter_worst);
area_worst = log(area_worst);
smoothness_worst = log(smoothness_worst);
compactness_worst = log(compactness_worst);
concavity_worst = log(concavity_worst);
concave_points_worst = log(concave_points_worst);
symmetry_worst = log(symmetry_worst);
fractal_dimension_worst =
log(fractal_dimension_worst);
run;

/* Plot the data */
ods graphics;
ods layout gridded columns=10 advance=table rows= 6;
proc univariate data= Cancer2 noprint;
hist radius_mean texture_mean perimeter_mean
area_mean smoothness_mean compactness_mean
     concavity_mean concave_points_mean
symmetry_mean fractal_dimension_mean radius_se
     texture_se perimeter_se area_se smoothness_se
compactness_se concavity_se
     concave_points_se symmetry_se
fractal_dimension_se radius_worst texture_worst
     perimeter_worst area_worst smoothness_worst
compactness_worst concavity_worst
     concave_points_worst symmetry_worst
fractal_dimension_worst;
run;
proc univariate data= Cancer2 noprint;
qqplot radius_mean texture_mean perimeter_mean
area_mean smoothness_mean compactness_mean
     concavity_mean concave_points_mean
symmetry_mean fractal_dimension_mean radius_se
     texture_se perimeter_se area_se smoothness_se
compactness_se concavity_se
     concave_points_se symmetry_se
fractal_dimension_se radius_worst texture_worst
     perimeter_worst area_worst smoothness_worst
compactness_worst concavity_worst
     concave_points_worst symmetry_worst
fractal_dimension_worst / normal(mu=est sigma=est
color=red l=2);
run;
ods layout end;


/* Standardized PCA */
proc princomp plots=all data=Cancer2 out=pca;
      var radius_mean texture_mean perimeter_mean
area_mean smoothness_mean compactness_mean
     concavity_mean concave_points_mean
symmetry_mean fractal_dimension_mean radius_se
     texture_se perimeter_se area_se smoothness_se
compactness_se concavity_se
     concave_points_se symmetry_se
fractal_dimension_se radius_worst texture_worst
     perimeter_worst area_worst smoothness_worst
compactness_worst concavity_worst
     concave_points_worst symmetry_worst
fractal_dimension_worst;
run;
proc print data= pca; run;

/* multivariate normal distribution for explanatory
variables */
proc sgscatter data = pca;
matrix prin1 prin2 prin3 prin4 prin5 prin6 / ellipse
=(alpha=.05);
run;

/* Logistic regression using pca */
proc logistic data = pca plots=all;
class  Diagnosis / param = ref;
model Diagnosis(event = "M") = prin1 prin2 prin3
prin4 prin5 prin6 / lackfit ctable;
output out=Probs predprobs= i lower=lcl upper=ucl;
run;
proc print data= Probs;
title 'Predicted Probabilities and 95% Confidence
Limits';
run;
```

VIII. VARIABLES

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)
The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.
All feature values are recoded with four significant digits.

Class distribution: 357 benign, 212 malignant

###################
Id - ID number
Diagnosis - The diagnosis of breast tissues (M = malignant, B = benign)
radius_mean - mean of distances from center to points on the perimeter
texture_mean -standard deviation of gray-scale values
perimeter_mean - mean size of the core tumor
area_mean
smoothness_mean - mean of local variation in radius lengths
compactness_mean - mean of perimeter^2 / area - 1.0
concavity_mean - mean of severity of concave portions of the contour
concave points_mean - mean for number of concave portions of the contour
symmetry_mean
fractal_dimension_mean - mean for "coastline approximation" - 1
radius_se - standard error for the mean of distances from center to points on the perimeter
texture_se - standard error for standard deviation of gray-scale values
perimeter_se
area_se
smoothness_se - standard error for local variation in radius lengths
compactness_se - standard error for perimeter^2 / area - 1.0
concavity_se - standard error for severity of concave portions of the contour
concave points_se - standard error for number of concave portions of the contour
symmetry_se
fractal_dimension_se - standard error for "coastline approximation" - 1
radius_worst - "worst" or largest mean value for mean of distances from center to points on the perimeter
texture_worst - "worst" or largest mean value for standard deviation of gray-scale values
perimeter_worst
area_worst
smoothness_worst - "worst" or largest mean value for local variation in radius lengths
compactness_worst - "worst" or largest mean value for perimeter^2 / area - 1.0
concavity_worst - "worst" or largest mean value for severity of concave portions of the contour
concave points_worst - "worst" or largest mean value for number of concave portions of the contour
symmetry_worst
fractal_dimension_worst - "worst" or largest mean value for "coastline approximation" - 1