Final project

Yuhan Wang

An Implementation of Non-parametric Logistic Regression for Breast Cancer Diagnosis and Comparison with Global Logistic Regression and Other Nonparametric Classification Approaches

Abstract

For medical data with binary outcome, logistical regression model is often used for binary classification. However, this parametric approach highly depends on the correctness of the pre-defined model and may not full capture the local pattern of the real-world dataset. In order to further explore this problem, several nonparametric models are implemented on Breast Cancer Wisconsin (Diagnostic) Data Set: additive logistic regression mode, k-nearest neighborhood method and conditional inference trees. ROC curves are drawn to compare the accuracy of outcome prediction between the parametric method and several nonparametric methods.

Introduction

Breast cancer is the most common cancer among women and one of the major causes of death among women worldwide. Detection at its early stages is highly important for the cancer to be treated effectively. Although there are many imaging techniques for cancer diagnosis, the accuracy achieved in each individual varies dramatically due to the subjectivity of visual diagnosis. The subjectivity which is inherent in visual diagnosis can be minimized with computational interpretation via statistical analysis. Logistic models are also used in risk screening and estimating the class probabilities of cancer type. However, it can only model the linear pattern of data. Therefore, in this project, generalized additive model is applied to add more flexibility into traditional logistic model and identify and characterize nonlinear regression effects.

Method

In this project, we use Breast Cancer Wisconsin (Diagnostic) Data Set, with 569 instances and 32 attributes. The attributes include subject id, one categorical variable of cancer diagnosis (357 benign, 212 malignant), 30 real-valued variables related with the cell nucleus characteristics, which are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The outcome we want to predict is cancer diagnosis, which is a binary variable, coded as 1 for benign diagnosis, 0 for malignant. 30 real-valued variables of cell nucleus characteristics are predictors.

To gain an unbiased evaluation of the model fit, the complete dataset was separated into training set and test set. 100 samples were randomly selected from original dataset as test set and the rest as training set. Several models were fit on training set and assess their accuracy on test set.

For the parametric approach, a logistic regression model is fitted on the training set with 30 real-value variables as predictors and cancer diagnosis as outcome. The fitted model is assessed on the test set by generating confusion matrix and ROC curve.
For additive logistic regression model, each linear term is replaced by a more general functional form

$$\log\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)} = a + f_1(X_1) + \cdots + f_p(X_p)$$

where each f_j is an smooth function. While the nonparametric form for the functions f_j makes the model more flexible, the additivity is retained and allows us to interpret the model in much the same way as before. In this case, natural cubic spline basis is applied on the first three covariates of the regression model, while other covariates remain linear. The optimal number of knots of natural cubic spline k=4 is chosen by 5-fold cross validation (see Figure 1). The additive logistic regression model is fitted on training set and then assessed on test set by generating confusion matrix and ROC curve.

For K-nearest neighborhood method (KNN), an empirically optimal k=17 was selected by 5-fold cross validation (see Figure 3).

For Conditional Inference Tree, this approach corrects for multiple testing to avoid overfitting issues, results in unbiased predictor selection and does not require pruning. Besides, decision tree could be useful when modeling human decisions, which can give precise thresholds for cancer diagnosis (see Figure 4). The conditional inference tree model is fitted on training data and assessed on test set by generating confusion matrix and ROC curve.

Result

By comparing AUC scores of ROC curves, we can see that on Breast Cancer Wisconsin (Diagnostic) Data Set, non-parametric logistic regression model performs the best (0.984), KNN the second (0.982), conditional inference tree the third (0.958), the parametric logistic regression model the last (0.955) (see Figure 5). By comparing confusion matrix from threshold=0.9 and prevalence=0.17, non-parametric logistic regression model has the highest accuracy (0.97), the parametric logistic regression model (0.96) the second, conditional inference tree the third (0.93), KNN the last (0.88) (see Table 1)

Table and Figure

Figure 1: choose the number of knots k=4 in natural cubic spline basis by 5- fold cross validation

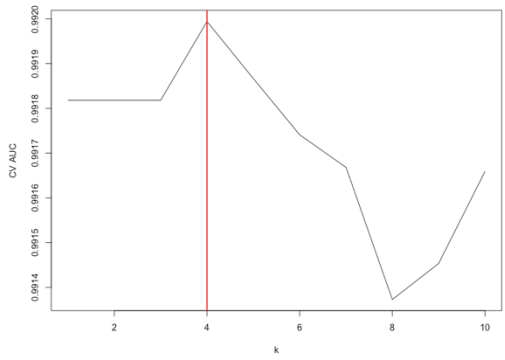

Figure 2: visualization the smooth functions of first nine covariates in additive logistic regression model

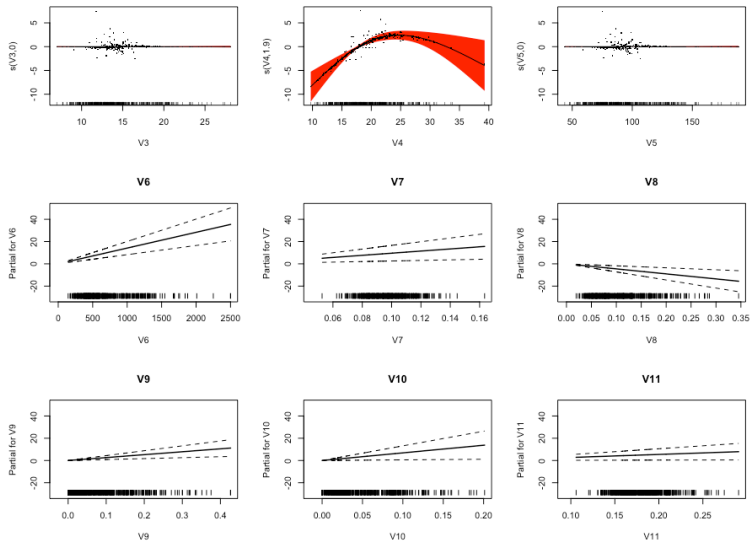

Figure 3: choose k=17 in K-nearest neighborhood method by 5- fold cross validation

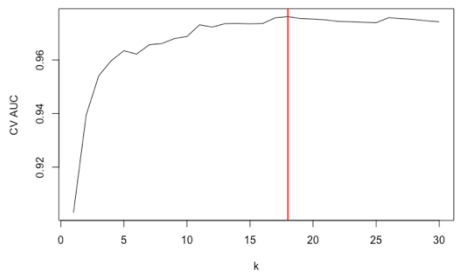

Figure 4: visualization of conditional inference tree on training set

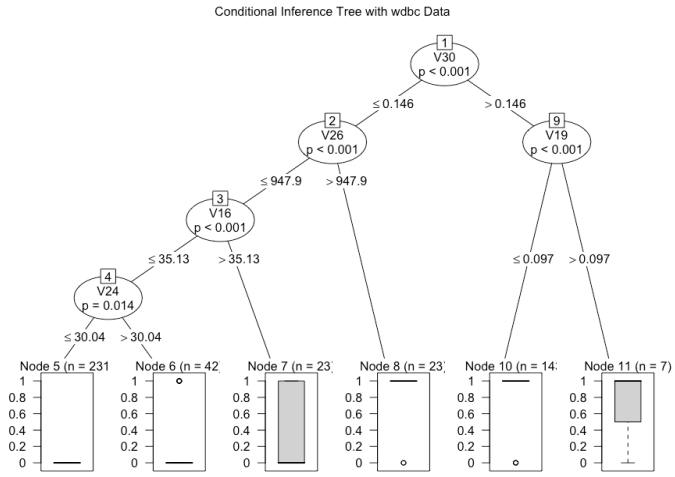

Figure 5: ROC Curve comparison on test set

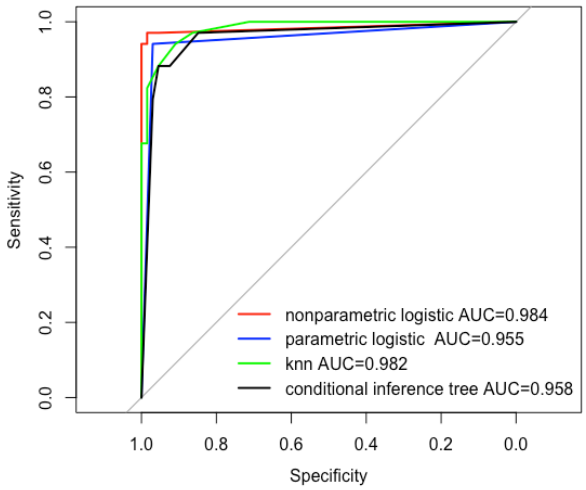

Table 1: comparison of confusion matrixes on test set

Additive logistic (accuracy=0.97)		truth	
		0	1
predict	0	64	1
	1	1	33

Logistic regression (accuracy=0.96)		truth	
		0	1
predict	0	64	2
	1	2	32

KNN (accuracy=0.88)		truth	
predict	0	66	12
	1	0	22

Conditional inference tree (accuracy=0.93)		truth	
		0	1
predict	0	63	4
	1	3	30

Discussion

The additive model is a useful extension of linear models with applying basis function, making linear model more flexible while still retaining much of their interpretability. However, there are also limitations in this project. First, the choice of covariates for basis function is arbitrary. More exploratory analysis need to be done to extract the non-linearity in the relationship between covariates and outcome. Second, the additive model in this project did not include interaction terms, which may not fully capture the pattern of data. Further study is needed to employ lasso-type penalties to estimate sparse additive models