Item reduction in a scale using constrained lasso method in logistic regression

Shaoxuan Chen

Abstract

In mental health studies, an underlying trait of interest is often not directly observable and scales are commonly used to measure it. The scales are composed of a set of items that are correlated to the latent trait. Excluding uninformative items from the scale will result in a reduced scale and it may maintain or even improve the prediction accuracy and classification accuracy of the full scale. In this project, constrained lasso method was performed in logistic regression to do item reduction among 40 olfactory tests. Additionally, evaluation based on Receiver Operating Characteristic (ROC) curve is applied to logistic regression and shows that the model with reduced scale improves both the prediction accuracy and classification accuracy of the model with full scale.

1 Introduction

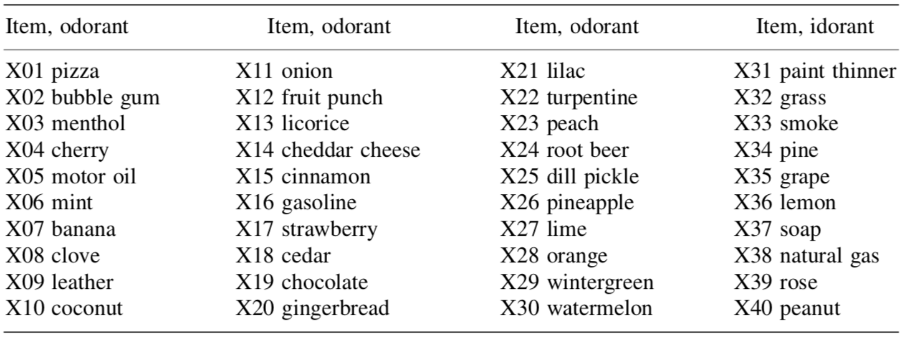
Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules. The outcome of a scaling process is a scale composed of a set of items that are correlated to the latent trait, and it can be unidimensional or multidimensional. The scale score, often defined as a sum of the response indicators of the items, is usually used as a measure of the latent trait. One rationale for the use of scale score is that the sum of item responses is a sufficient statistic for the underlying latent trait. The ideal case is that all items in a scale are consistent, highly correlated with latent variable, less correlated to each other and items have high classification ability.[1] If the ideal case is not satisfied, e.g., the scale is composed of some uninformative items which are not useful in classification, the "noisy" items should be checked and removed from the scale. Excluding uninformative items from the scale will result in a reduced scale, and reduced scale may maintain or even improve the classification accuracy of the full scale. Compared with full scale, using the reduced scale to test patients and do predictions will be more efficient and cost-effective.[1]

In mental health studies, an underlying trait of interest is often not directly observable and scales are commonly used to measure it. Since unidimensional scale is useful in screening for individuals at risk of a certain illness if it measures the underlying latent trait of the illness, in this project we just perform analysis on unidimensional scale.

The background information of this project is that the development of Alzheimer' s disease (AD) is associated with olfactory identification defect. That is to say, if patients have poor olfactory identification ability, they are more likely to develop AD. The deficits are often measure by the standardized University of Pennsylvania Smell Identification Test (UPSIT), which is a self-administered 40-item scratch-and-sniff multiple choice odor identification test consists of four booklets each containing 10 odors, with one odor per page. Subjects are instructed to scratch the label, then sniff the label and choose a response category closest to the smell that they experienced. [1] The 40 olfactory tests are shown in Figure 1.

The data source of this project is Olfaction test data collected from 127 patients with mild cognitive impairment (MCI) who were at risk to develop AD. The 127 patients were administered UPSIT, a 40-

Figure 1: 40 olfactory test(UPSIT)



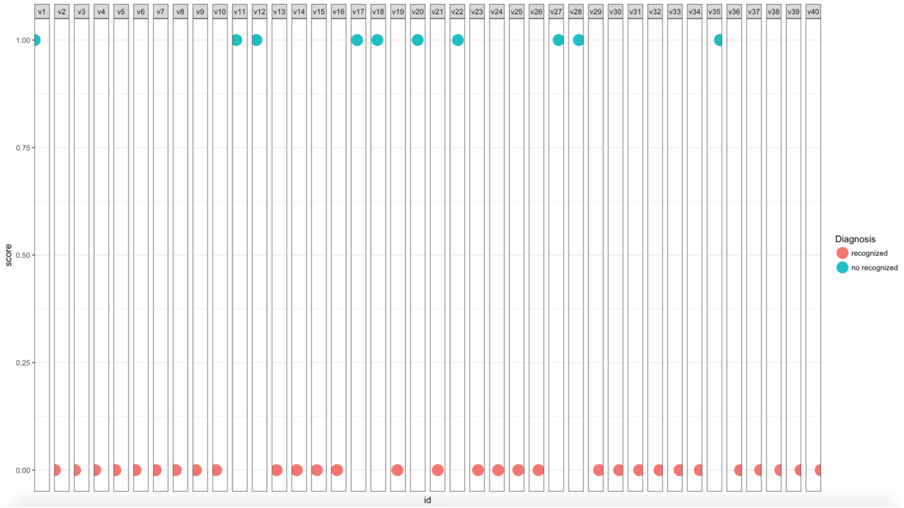
Item, odorant	Item, odorant	Item, odorant	Item, idorant
X01 pizza	X11 onion	X21 lilac	X31 paint thinner
X02 bubble gum	X12 fruit punch	X22 turpentine	X32 grass
X03 menthol	X13 licorice	X23 peach	X33 smoke
X04 cherry	X14 cheddar cheese	X24 root beer	X34 pine
X05 motor oil	X15 cinnamon	X25 dill pickle	X35 grape
X06 mint	X16 gasoline	X26 pineapple	X36 lemon
X07 banana	X17 strawberry	X27 lime	X37 soap
X08 clove	X18 cedar	X28 orange	X38 natural gas
X09 leather	X19 chocolate	X29 wintergreen	X39 rose
X10 coconut	X20 gingerbread	X30 watermelon	X40 peanut

item olfactory test, at baseline. And they were followed for at least two years. There were 31 patients who met the criteria of AD diagnosis within two years after baseline evaluation, being considered in the group with high risk of developing AD (D =1). The low-risk group (D =0) had 96 patients who did not develop AD in the two years of follow-up.[2]. The data set contains 127 entries(127 patients) and 41 variables (1 response variable indicate whether patients are likely to develop AD after two years of follow up; 40 variables with binary responses represent the odour items). The data set was divided into training set(%70, about 100 patients) and testing set (30%, about 27 patients) to do analysis in later parts.

2 Exploratory data Analysis

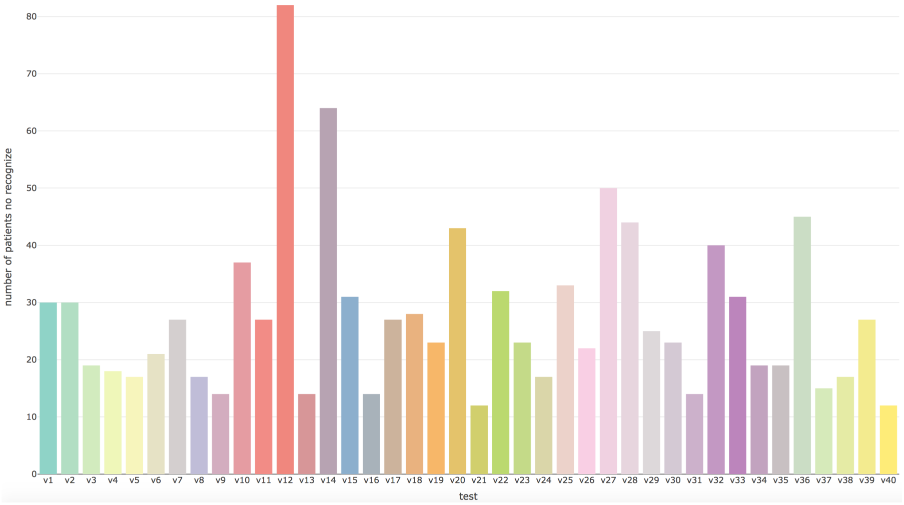
If patients didn't recognize the odor, he/she would get one point in the test. Figure 2 shows the performance of the first patient. It's obvious that the first patient didn't recognize 10 odors among the 40 olfactory tests.

Figure 2: performance of the first patients in 40 olfactory tests



By comparing the items in reduced scale obtained from constrained lasso method, we can check whether the number of patients fail in each odor test has direct relationship with the latent interest of this study. Figure 3 shows the plot of total number of patients who didn't recognize the odor in each test.

Figure 3: number of patients fail in each odor test



3 Statistical method

3.1 Logistic regression

In classification, when the item responses are binary variables, logistic regression assumes normality for quantitative classifiers [3]. In logistic regression, logit link is used between probability of disease class and a linear combination of predictors. In this project, the disease class is whether patients being considered in the group with high risk of developing AD (D =1) or with low risk of developing AD (D =0). However, in 2001, Hastie et al. [4] showed that it is hard to verify regression models, especially in high dimensions. And in 2006, Pepe et al. [5] also pointed out that logistic regression model might have poor performance in classification. Therefore, to make sure the fitted regression model is reasonable and have good performance, we should prune the data to make the scale close to unidimensional and items in the scale are equally weighed.

3.2 Factor analysis & Latent variables

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. It searches for joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms; thus, the factors can also be treated as the variables in regression analysis. Factor loading is the regression coefficient of a factor in predicting an item. It can also be treated as the parameter estimates. However, we know from the factor analysis that some of the items had greater factor loadings than the other ones comprising that scale, or some of the items contain higher-dimensional information. They are thus explaining more of the variance. Using those factor loadings is possible to give unequal weights to items, which, in this project, doesn' t meet the criterion of equal weighed items required in logistic regression. Therefore, we should do item reduction by constrained the coefficients to get unidimensional scale with equal weighed items.

3.3 Constrained lasso (Positive lasso)

Lasso:

$$\arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Positive Lasso:

$$\arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1, \textit{subject to } \beta \geq 0 \text{ and } \|\beta\| \leq \lambda$$

Lasso, with L1 penalty, the advantage is that the solutions are typically sparse, which can be used for variable selection, when $\lambda \longrightarrow 0$,we have the least squares solution; as $\lambda \longrightarrow \infty$, the solution approach 0. As for Positive lasso, it requires the lasso coefficients to be non-negative and it's a part of constrained lasso. Because of the relationship between regression analysis and factor analysis, by constrained the coefficients, we can constrain the factor loadings and prune the items of the full scale. In this way, by using positive lasso we are able to obtain the scale with reduced items which is unidimensional.

3.4 Receiver Operating Characteristic (ROC) curve & Area Under the ROC curve (AUC)

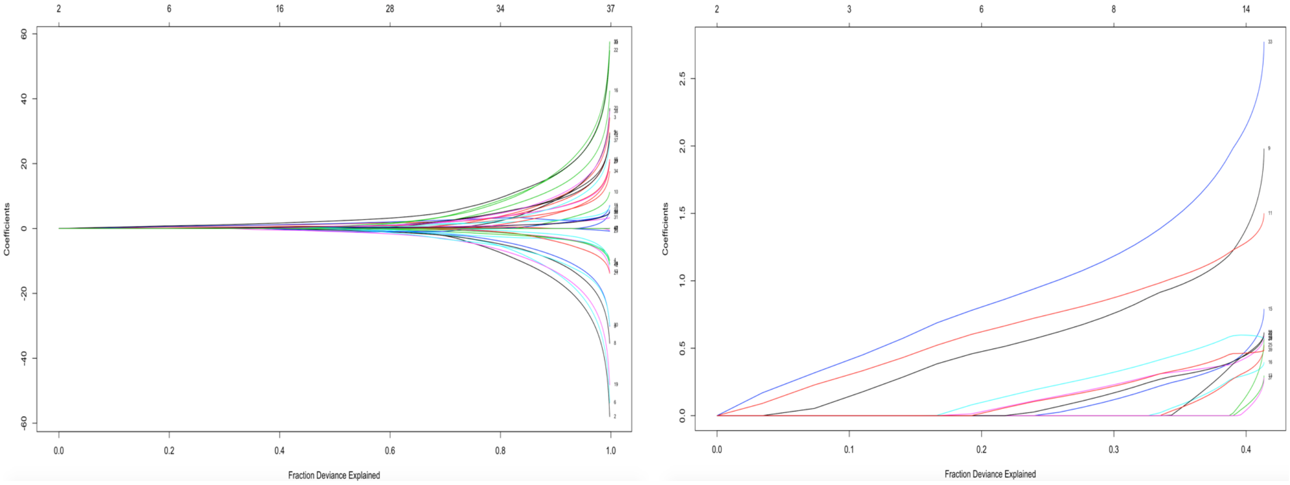
Receiver Operating Characteristic (ROC) curve is widely accepted as a standard diagnostic technique to assess the performance of classifier. It shows how sensitivities change either with specificities or with false positive proportions(1-specificities) for all possible cutoff scores. The Area Under the Curve (AUC) is a summary of the ROC curve. Its a common metric to represent the power of classifier. In this study, the classification accuracy is defined similarly to AUC. Based on the evaluation of the change in classification accuracy due to inclusion or exclusion of an item, we select items for a reduced unidimensional scale.

4 Result

4.1 Change of coefficients against with deviance/L1-norm/Log-lambda in unconstrained/constrained model

In the following three coefficient plots, each colored line in the plot represents the coefficient trend of one predictor in the regression model (or one item in the scale). The numbers above the upper line of the plot represent the number of predictors in the regression model.

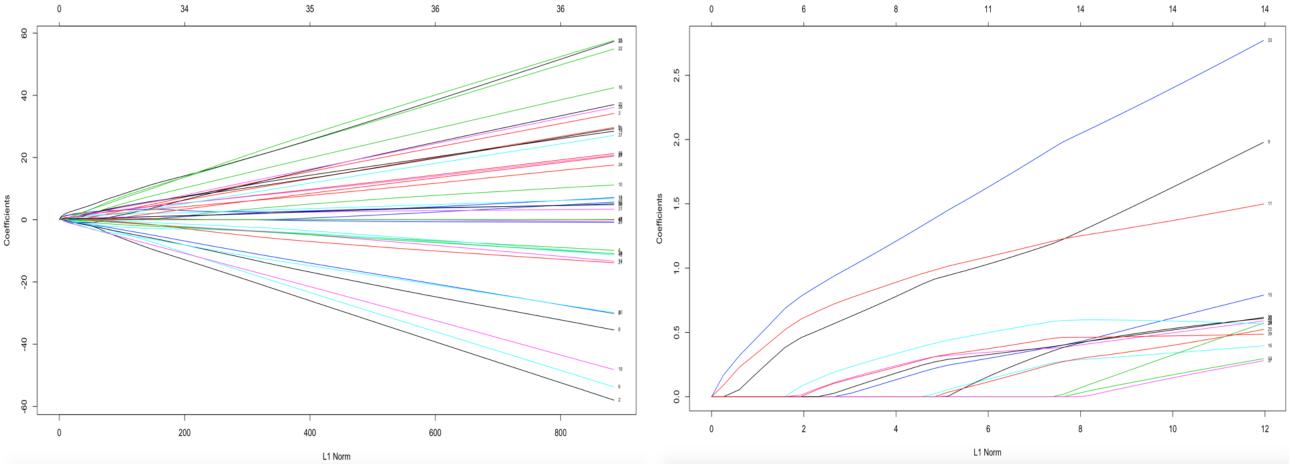
Figure 4: Change of coefficients against with deviance (left: unconstrained model, right:constrained model)



Deviance is a measure of goodness of fit in GLM. It can be used to perform model comparison for it provides a measure of goodness-of-fit of the model being evaluated when compared to the null model.

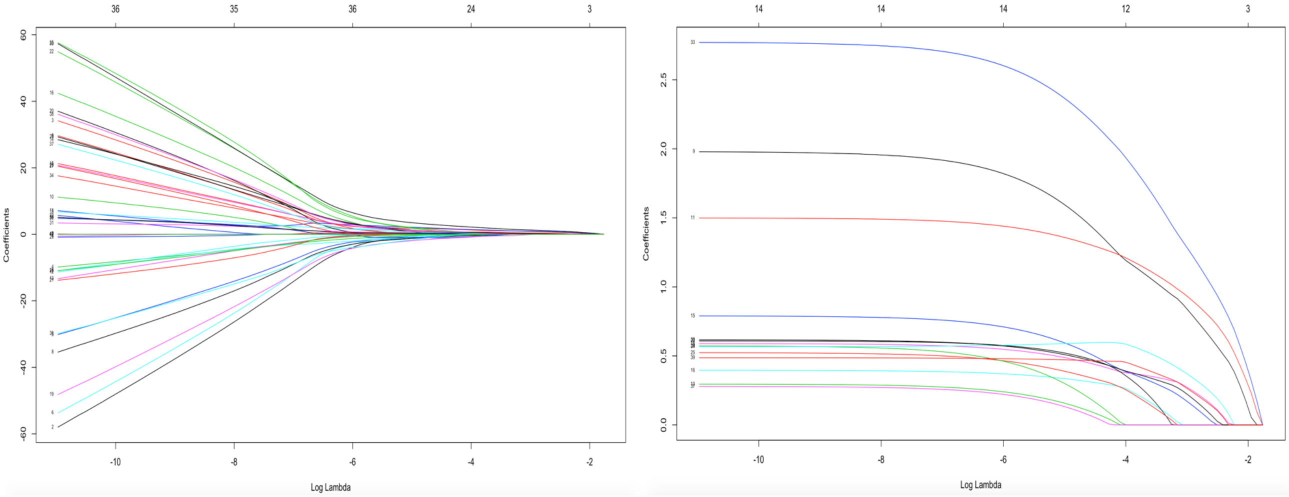
From Figure 4, in unconstrained model, with the number of predictors in the model increase from 2 to 37, the corresponding fraction deviance explained by the model with reduced scale increasing from 0 to 1 and the absolute value of coefficients increase from 0 to 60. As for constrained model, compared with the unconstrained model, the values of coefficients were constrained to all positive. The number of predictors shrinks from 2-37 to 2-15. The corresponding fraction deviance explained by the model with reduced scale shrinks from 0-1 to 0-0.43, and the range of coefficient shrinks from 0-60 to 0-3.

Figure 5: Change of coefficients against with L1-norm (left: unconstrained model, right:constrained model)



In lasso regression we use L1-norm penalty term. The L1-norm of a vector is defined as the sum of the absolute values of its components. From Figure 5, in unconstrained model, with the number of predictors in the model increases from 0-40, the corresponding L1-norm of model with reduced scale increase from 0 to 900 and the absolute value of coefficients increase from 0 to 60. As for constrained model, compared with the unconstrained model, the values of coefficients were constrained to all positive. The number of predictors shrinks from 0-40 to 0-14. The corresponding L1-norm of model with reduced scale shrinks from 0-900 to 0-12, and the range of coefficient shrinks from 0-60 to 0-3.

Figure 6: Change of coefficients against with Log-lambda (left: unconstrained model, right:constrained model)



λ is the weight given to the penalty term (the L1 norm), so as λ grows, the penalty term has greater effect and will result in fewer variables in the model (because more and more coefficients will be zero-valued). From Figure 6, in unconstrained model, with the log-λ increase from -12 to -2, i.e., more weight put on the penalty term, the corresponding number of predictors in the model decreases from 40 to 0 and the absolute value of coefficients decrease from 60 to 0. As for constrained model, compared with the unconstrained model, the values of coefficients were constrained to all positive. As the log-λ increase from -12 to -2, the corresponding number of predictors shrinks from 0-40 to 3-14. The range of coefficient shrinks from 0-60 to 0-3.

4.2 Result of cross validation and corresponded items in reduced scale

The result of cross validation performed on the training data shows that λ_{min}=0.08156105, which corresponding to 8 predictors in the logistic regression model and it means that the reduced scale is composed of 8 items. The predictors and coefficients are shown in Table 1.

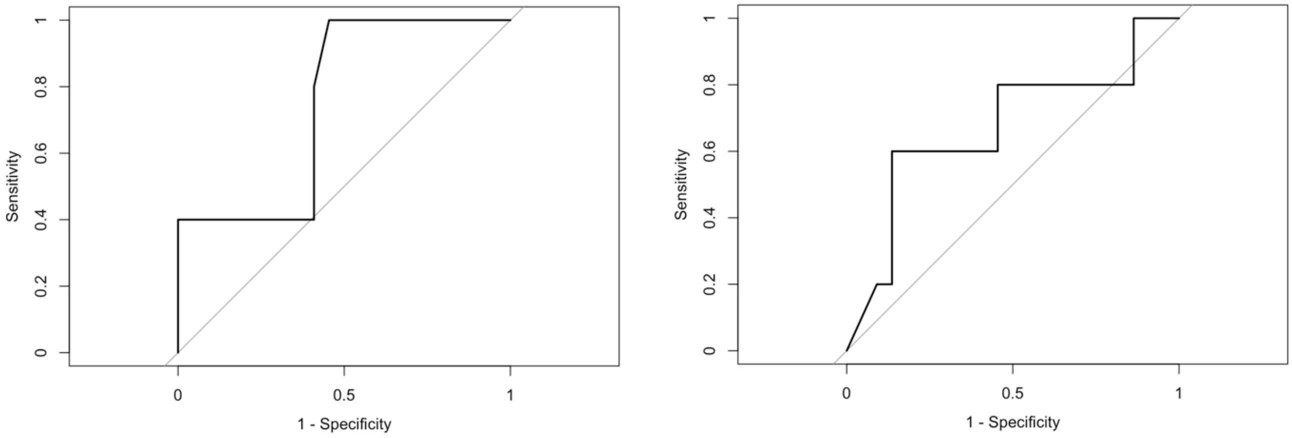
predictor	coefficient
(intercept)	-2.14496674
V8	0.72631761
V10	0.85010231
V14	0.09650135
V17	0.21071366
V32	1.14143327
V33	0.21071366
V37	0.14613522
V38	0.19928347

Table 1: Predictors and corresponded coefficients of model with reduced scale

4.3 Comparison of full scale/reduced scale logistic regression model

In this project, ROC curve method was performed on the testing dataset to assess the power of classifiers. And in this study, the classification accuracy is defined similarly to AUC. Figure 7 shows the ROC curves of model with reduced scale and model with full scale. The prediction accuracy and classification accuracy of model with reduced scale and model with full scale are also summarized in Table 2.

Figure 7: ROC curves of model with reduced/full scale



	prediction accuracy	classification accuracy
full scale	0.7778	0.6727
reduced scale	0.8519	0.7500

Table 2: Prediction accuracy and classification accuracy of model with reduced/full scale

5 Discussion

The constrained lasso method suggests inclusion of 8 items for a reduced scale R8 = X8, X10, X14, X17, X32, X33, X37, X38. It means compared with performing 40 olfactory tests, using 8 odor items: clove, coconut, cheddar cheese, strawberry, grass, smoke, soap as well as natural gas to do the olfactory test is even more efficient and cost-effective. The estimated classification accuracy A(R8) was 0.75, exceeding the 0.6727 of the full scale. The estimated prediction accuracy P(R8) is 0.8519, also exceeding 0.7778 of the full scale. Therefore, we can conclude that the logistic regression model with reduced scale can improve both the classification accuracy and the prediction accuracy of the model with full scale.

Compare the information provided in the exploratory data analysis part with the reduced items that we chose, we can draw the conclusion that the number of patents fail in each test is not correlated with the reduced items. That is to say, whether the odor is difficult/easy to be recognized has no direct relationship with the latent interest of this study.

Define π_i, the probability of people being considered in the group with high risk of developing AD, the final logistic regression model is :

$$log\frac{\pi_i}{1-\pi_i}=-2.15+0.73\times I_{v8=1}+0.73\times I_{v8=1}+0.85\times I_{v10=1}+0.1\times I_{v14=1}+0.21\times I_{v17=1}+1.14\times I_{v32=1}+0.3\times I_{v33=1}+0.15\times I_{v37=1}+0.2\times I_{v38=1}$$

References

[1] Liu, X. and Jin, Z. (2007). Item reduction in a scale for screening, Statistics in Medicine, 26, 4311-4327.

[2] Doty RL, Shaman P, Dann M. Development of the University of Pennsylvania Smell Identification Test: a standardized microencapsulated test of olfactory function. Physiology Behavior 1984; 32:489502.

[3] Press SJ, Wilson S. Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association 1978; 73:699705.

[4] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer: New York, 2001.

[5] Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. Biometrics 2006; 1:221229.