Assignment Learning from Data

0501050 Ting-Shuo Yo
0501247 Shriprakash Sinha

1 Introduction

Given a dataset of size $n = 200$, a classification model supposed to be constructed for the underlying population. The dataset contains numerical independent variables $\mathbf{X} = \{X_1, X_2, X_3\}$ and binary class label Y, and not all independent variables are predictive for Y.

To learn a classifier with minimal error rate and to provide an proper estimate for the error rate, a systematic analysis process is performed. The rationale of this analysis process is discussed in section 2, and the corresponding results of each step are summarised in section 3. In section 4, we give the best learned model and a few concluding remarks.

2 Methods

Following is the analysis procedure used for this assignment:

1. **Explore the given dataset**
   First step of the analysis is to explore the basic properties of the given dataset. Since it is hinted that some independent variables are not predictive at all, to find a proper formula for models to fit may be a good start. In this step, we look at the correlation coefficients among variables and perform a few runs of step-wise logistic regression. Afterward, a few candidate formulas are used for later steps.

2. **Test several classification algorithms**
   Following classification algorithms are tested: logistic regression, linear/quadratic discriminant analysis, neural networks, k-nearest neighbour, and support vector machines. Each classifier is tested with a few combination of parameters, and the most proper parameter sets are decided upon the resulting error rates of the re sampling tests.

   With the "best" parameter sets, all algorithms are compared together by their error rates on the same re sampling datasets. A few candidate algorithm-parameter combination are used for the next step.

3. **Bootstrap aggregating with adaptive resampling**
   For all survival candidates in previous step, bagging and boosting techniques are applied to improve their performance (i.e., to lower the error rates). **The final decision will be made upon the final performance measure.**

3 Results

Candidate formulas

The correlation matrix for $x.1, x.2, x.3$, and y shows that $x.3$ has little correlation to other variables (< 0.1). Meanwhile, the step-wise logistic regression (use AIC for model selection) with 20-fold cross validation also shows a preference to the formula: $y = x.1 + x.2^2$ (17 out of 20). Therefore, two formulas, i.e., linear and $y = x.1 + x.2^2$ are considered for following tests, and are referred as f1 and f2.

Candidate classifiers and parameter sets

The error rates are used as the performance measure, and calculated based on in-sample, cross-validation (2-, 5-, 10-, 20-, 50-, 100-fold, and leave-one-out) and bootstrapping (100 resampling). The best parameter set for each classifier is described below.

1. **logistic regression**
   For in-sample and cross-validation with $2 \sim 100$ folds, models with f1 perform slightly better than ones with f2, as well as the estimation with bootstrapping. However, for leave-one-out cross-validation, f1 and f2 are with similar error rate. Hence, both models are considered for next step, and are referred as LR1 and LR2.

2. **linear/quadratic discriminant analysis**
   For all error rate estimates, f2 is slightly better. Also, LDA performs slightly better than QDA. Here, we decide to use both LDA and QDA with f1 for next step.

3. **k-nearest neighbour**
   No formula is applied to this non-parametric classifier. The value of k is tests from 1 to 10, and the best performance can be found around $k = 5 \sim 10$. This does not include the in-sample test, because the 1NN will always be 100% correct in this case. Therefore, $k = 5$ is used for later step.

4. **neural network**
   An implementation of one-hidden-layer neural network is tested with formula $= \{f1, f2\}$, size of hidden layer $= \{2, 4, 6, 8, 10\}$, decay $= \{0.01, 0.001, 0.0001\}$, and activation functions $= \{softmax, entropy\}$ ($softmax$ is tested only for f1 due to technical reasons). After comparing the estimated error rates, $\{size = 4, decay = 0.01, entropy\}$ for both f1 and f2 are selected.

5. **support vector machine**
   A SVM implementation in the package **e1071** is used for the test. Kernels can be chosen from $\{linear, polynomial, radial - basis, sigmoid\}$, and $linear$ kernel is selected along with formula f1.

A summary of the tests is shown in table 1. Since the 5-NN tends to have higher error, it will not be considered in the next step.

Bagging and Boosting

All classifiers in table 1 except 5-NN and SVM are applied to bootstrap aggregating and arc-x4 of 10, 50, and 100 models trained by resampling. The error rate of ensemble predictions are evaluated by 10-fold cross-validation and are summarised in table 2. By comparing the performance in previous experiments with one single model, we mark those get improvements from ensemble prediction with a *.

Table 1: Summary of the mean error rates for each classifier.

	LR1	LR2	LDA1	QDA1	5-NN	NNET1	NNET2	SVM
in-sample	0.250	0.245	0.250	0.240	0.200	0.140	0.116	0.240
10-fold CV	0.250	0.255	0.260	0.270	0.265	0.270	0.235	0.270
leave-one-out	0.260	0.260	0.265	0.275	0.290	0.285	0.255	0.260
100 bootstrap	0.265	0.259	0.260	0.276	0.311	0.265	0.265	0.263

Table 2: Error rates (average of 10-fold CV) of bagging and arc-x4 for candidate classifiers. Classifiers get improved performance compared to one single model are marked with a *.

	LR1	LR2	LDA1	LDA2	QDA1	QDA2	NNET1	NNET2
10 BAG	0.255	0.260	0.260	0.250	0.255*	0.270	0.280	0.225*
50 BAG	0.255	0.260	0.270	0.250	0.250*	0.270	0.275	0.265
100 BAG	0.255	0.260	0.265	0.250	0.250*	0.270	0.295	0.255
10 ARC	0.295	0.245*	0.280	0.265	0.250*	0.265	0.330	0.260
50 ARC	0.260	0.250*	0.235*	0.260	0.255*	0.250*	0.295	0.280
100 ARC	0.265	0.245*	0.255*	0.250	0.255*	0.260*	0.285	0.260

As shown in table 2, QDA algorithm benefits from all kinds of ensemble, while LR2, LDA1, and NNET2 get improved for some cases. However, if we further choose the top two classifiers, i.e., LDA1-50-ARC and NNET2-10-BAG, for leave-one-out cross-validation, their estimated error rate will both be 0.265, which is not better than the performance of one single model. The final decision will be discussed in next section.

4 Concluding remarks

Table 1 shows the average error rate for each of the classifiers using different techniques for model selection and assessment. Looking column-wise and considering all the classifiers, we would prefer LDA1 to be our best model. The reason for selecting LDA1 is its stability over mean error rates across the various techniques. Since the mean error rates are almost same for LDA1, LR1 and LR2, we would choose LDA1. In Table 2 the techniques of model selection and assessment use mixtures of bagging, CV and arcing.

In contrast to LDA1 which was chosen from Table 1, we may choose QDA1 here as the best model. This choice is based on the observation of somewhat stable low generalization error results across the mixture of techniques used. LDA1 does not perform better in case of bagging but it does perform better in case of Arcing. Thus in conclusion we may choose, QDA1 as the model, with bagging with 100 runs and 10-fold CV, as the method of model selection and assessment. For this case of data set, we can also say that mixture of techniques like Bagging and CV may perform better than the techniques considered individually.

Regarding the statistics conserning the 10-fold CV for QDA1-100-BAG, we have a mean error rate of $\mu = 0.25$, and standard deviation 0.085. With 95% confidence interval and degree of freedom = 9 (because we have 10 different error rates), the *t-score* is $t(df = 9, p = 0.05) = 2.26$. The estimated standard deviation of the population, is then .0283. Thus with 95% confidence we can state that the estimated error rate lies in the interval $[0.186 \sim 0.314]$ with the mean value $\mu = 0.25$.