Common Classification Methods Study

Written report for STAT:7400 Computer Intensive Statistics

by Yiheng Liu

1 Introduction

With the development of machine learning and statistics, a lot of classification methods have been developed. On the basis of a training set of data containing observations whose category membership is known, classification is to identify which of a set of categories a new observation belongs to. Because each method has its own characteristics and is more efficient when dealing with some particular type of data sets, questions naturally arise: What are the advantages and disadvantages of these methods? How well do they perform on different kinds of data sets? Therefore, this project is proposed to implement and study some common classification methods in R.

For the first part, a brief introduction is given to three basic classification methods: Logistic Regression (LR), Linear Discriminative Analysis (LDA) and Naive Bayes (NB). Then I developed a R package to implement these three method and carry out a simulation study with 5 different data scenarios.

2 Basic Classification Methods

In this section, a brief introduction is given to LR, LDA and NB. To simplify the study, the observations are assumed to fall into two categories only. For convenience, I am using the generic 0/1 coding for the response Y. Suppose there are p predictors and denote them by $X = (X_1, X_2, ..., X_p)$.

2.1 Logistic Regression

Suppose $p(X) = \Pr(Y = 1|X)$, we want to model the relationship between $p(X)$ and X. It is given by

$$\log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p.$$

This can be rewritten as

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + ... + \beta_p X_p}}.$$

We use the maximum likelihood method to estimate $\beta_0, \beta_1, ..., \beta_p$. Usually, an observation is classified as 1 if $\hat{p}(X) \geq 0.5$.

2.2 Linear Discriminative Analysis

Compared to LR directly modeling $\Pr(Y = 1|X = x)$ using the logistic function, LDA models the distribution of X given Y and then use Bayes' theorem to flip the these around into estimates for $\Pr(Y = 1|X = x)$.

Let π_k represent the prior probability that an observation comes from the kth class (i.e. $\pi_k = \Pr(Y = k)$). Let $f_k(X) = \Pr(X = x|Y = k)$ denote the density of X that comes from the kth class. Then, by Bayes' Theorem,

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{2} \pi_l f_l(x)}.$$

Suppose we model each class density as multivariate Gaussian

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}.$$

In LDA, we also assume that $\Sigma = \Sigma_k \forall k$. When comparing two classes 0 and 1, it is sufficient to look at the log-ratio,

$$\log \frac{\Pr(Y = 1|X = x)}{\Pr(Y = 0|X = x)} = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}(\mu_1 + \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0) + x^T \Sigma^{-1}(\mu_1 - \mu_0).$$

In practice, we need to estimate the unknown parameters by our training data set.

- $\hat{\pi}_k = N_k/N$;
- $\hat{\mu}_k = \sum_{gi=k} x_i/N_k$;
- $\hat{\Sigma} = \sum_{k=0}^{1} \sum_{gi=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T/(N - 2)$;

The LDA rule classifies to class 1 if

$$x^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) > \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_0)^T \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) - \log(N_1/N_0)$$

and class 0 otherwise.

2.3 Naive Bayes

Similar to LDA, NB estimates $\Pr(Y = 1|X = x)$ by Bayes' Theorem. The difference is it assumes the predictors are conditionally independent for each class. Therefore, the joint distribution of predictors given each class can be estimated marginally. For categorical predictors, the marginal probability mass function given each class is estimated by counting measure. For continuous predictors, the marginal density function give each class is estimated by Gaussian.

Given a new observation X^{new}, the decision rule is

$$Y^{new} \leftarrow \arg\max_{y_k} \prod_i \Pr(X_i^{new}|Y = y_k) \prod_j N(X_j^{new}; \mu_{jk}, \sigma_{jk}).$$

3 Simulation Study

3.1 R package `class2`

To carry out a simulation study on these 3 methods, I developed a R package `class2` with two functions in it. `classify` is used to train the classifier given the training data set. Then we can use `classify.predict` to evaluate the performance of the classifier given the test data set.

3.2 Data Scenarios

Suppose there are only two predictors, I compared the performance of these classifiers based on five data scenarios. For each data scenario, a test data set of size 500 is generated in each class. Then I trained these classifiers on different training data set 1000 times. A training data set of size 20 in each class is generated each time. In the end, the mean error rate is reported for each classifier. Detailed information of each data scenario is given below.

1. $X_1|Y=0 \sim N(1,1)$; $X_2|Y=0 \sim N(2,4)$; $X_1|Y=1 \sim N(3,1)$; $X_2|Y=1 \sim N(4,4)$; $X_1 \perp X_2$
2. $X_1|Y=0 \sim N(1,1)$; $X_2|Y=0 \sim N(2,4)$; $X_1|Y=1 \sim N(3,1)$; $X_2|Y=1 \sim N(4,4)$; $Corr(X_1,X_2)=0.5$
3. $X_1|Y=0 \sim t(1)+1$; $X_2|Y=0 \sim t(2)+2$; $X_1|Y=1 \sim t(1)+3$; $X_2|Y=1 \sim t(2)+4$; $X_1 \perp X_2$
4. $X_1|Y=0 \sim N(0,1)$; $X_2|Y=0 \sim B(1,0.8)$; $X_1|Y=1 \sim N(2,1)$; $X_2|Y=1 \sim B(1,0.2)$; $X_1 \perp X_2$
5. $X_1|Y=0 \sim N(0,1)$; $X_2|Y=0 \sim B(1,0.8)$; $X_1|Y=1 \sim N(2,1)$; $X_2|Y=1 \sim B(1,0.2)$; $Corr(X_1,X_2)=-0.5$

3.3 Simulation Results

The error rate is given below for each combination of classifier and data scenario.

	1	2	3	4	5
LR	0.1696	0.1950	0.2344	0.1251	0.1373
LDA	0.1677	0.1940	0.2491	0.1311	0.1408
NB	0.1690	0.1948	0.3463	0.1231	0.1334

Table 1: Error rate of 5 Data Scenarios

1. For independent normal predictors, three classifiers have similar performance . LDA outperforms the other two since this is the model assumed by it.

2. For correlated normal predictors, their performance becomes a little bit worse due to the violation of assumptions.

3. For independent t distributions, LR outperforms the other two because it has a more general set-up.

4. For independent normal and bernoulli predictors, NB outperforms the other two because its expertise in dealing with categorical variables.

5. For correlated normal and bernoulli predictors, their performance becomes a little bit worse. But NB still outperforms the other two.

3.4 Conclusion

In summary, LR is a more general method in dealing with different kinds data set. But it has computationally complex iterative algorithm and may not converge. LDA is very easy to implement and performs well with normal assumptions. But it is not flexible when assumptions deviate from multivariate normal. NB has a good performance when dealing with mixed (categorical and continuous) data set. But it cannot perform well the predictors are correlated and not normal.

REFERENCES

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2015). An Introduction to Statistical Learning with Applications in R.