Classification and Discriminant Analysis: Bayes Classifiers and logistic regression for Home equity loan data

Mingliang Wang

1 Abstract

The consumer credit department of a bank wants to automate the decision-making process for approval of home equity lines of credit. To do this, they will follow the recommendations of the Equal Credit Opportunity Act to create an empirically derived and statistically sound credit scoring model. The model will be based on data collected from recent applicants granted credit through the current process of loan underwriting. The model will be built from predictive modeling tools, but the created model must be sufficiently interpretable to provide a reason for any adverse actions (rejections). In this work, we compared the performance of Bayes classifiers with logistic regression in the home equity data.

2 Bayes classifier

Suppose we have the cost function as

$$L(g_k, \hat{G}(x)) = \begin{cases} 0 & \hat{G}(x) = g_k \text{ or } \text{“ correct classification”} \\ 1 & \hat{G}(x) \neq g_k \text{ or } \text{“ miss classification”} \end{cases} \tag{1}$$

where g_k is the true label of x and $\hat{G}$ is the label estimated by a classifier. The cost of Bayes classifier can be written as

$$\min_{\hat{G}} L(g_k, \hat{G}(x)) \Pr(Y = g_k \mid X) \tag{2}$$

where X denotes the input features. The Bayes classifier weighs the cost of miss-classification by the conditional probability $\Pr(Y = g_k \mid X)$. The interpretation of this classifier is easy to understand: $\Pr(Y = g_k \mid X)$ is the evidence of the label being g_k, if the estimate $\hat{G}$ of a classifier (or a model) indicates otherwise, then the cost is violating the evidence under the model (or the classifier) which is $\Pr(Y = g_k \mid X)$. The Bayes classifier can be rewritten as

$$\min_{\hat{G}(x)} \sum_{g_k \neq \hat{G}(x)} \Pr(Y = g_k \mid X) \tag{3}$$

$$\Rightarrow \min_{\hat{G}(x)} 1 - \Pr(Y = \hat{G}(x) \mid X) \tag{4}$$

$$\Rightarrow \max_{\hat{G}(x)} \Pr(Y = \hat{G}(x) \mid X) \tag{5}$$

Therefore, the Bayes classifier is to choose the class with maximum probability.

2.1 Naive Bayes classifiers

The variables in X are said to be conditionally independent of Y, given Z, if the following condition holds:

$$\Pr(X, Y \mid Z) = \Pr(X \mid Z) \Pr(Y \mid Z) \tag{6}$$

Now, the naive Bayes assume the input features $X = \{X_1, \cdots, X_d\}$ are conditional independent,

$$\Pr(Y \mid X) = \frac{\Pr(Y) \prod_{i=1}^{d} \Pr(X_i \mid Y)}{\Pr(X)} \tag{7}$$

Then, the naive Bayes classifier choose the class that maximize the numerator term $\Pr(Y) \prod_{i=1}^{d} \Pr(X_i \mid Y)$. The estimation of $\Pr(X_i \mid Y)$ in naive Bayes classifier is given as follows regarding to the categorical and continuous attributes:

- for categorical attributes: $\Pr(X_i = x \mid Y = y) =$ The fraction of instants of x in class y.
- for continuous attributes, we can

1. discretize each continuous attribute and then replace the continuous attribute value with its corresponding discrete interval
2. assume a certain form of probability distribution for the continuous variable and estimate the parameters of the distribution using the training data. Most often, a Gaussian distribution is chosen.

$$\Pr(X_i = x_i \mid Y = y_j) = \mathcal{N}(\mu_{i,j}, \sigma_{i,j}^2)$$

2.2 Linear discriminant analysis

One of the most common ways to represent the classifiers is in terms of a set of discriminant functions $\delta_i(X),\ i = 1, \cdots, K$, where

$$\delta_i(X) = \Pr(\omega_i \mid X) \propto \Pr(X \mid \omega_i) \Pr(\omega_i) \text{ or} \tag{8}$$

$$\delta_i(X) = \ln \Pr(X \mid \omega_i) + \ln \Pr(\omega_i) \tag{9}$$

and K is the number of classes. We determine the label (or class) of a sample as $\{i \mid \delta_i(X) \geq \delta_j(X), j \neq i, j = 1, \cdots, K\}$.

If we assume the conditional probability $\Pr(X \mid \omega_i)$ is Gaussian distributed, then we have

$$\Pr(X = x \mid \omega_i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{10}$$

Therefore, the discriminant function can be rewritten as

$$\delta_i(X) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \ln \Pr(\omega_i) \tag{11}$$

Then the decision boundary between class k and i is

$$\delta_k(X) = \delta_i(X) \tag{12}$$

The decision boundary of a binary classification is shown in Figure 1. The probability distribution parameters can be easily estimated using the samples,

$$\hat{\pi}_k := \hat{\Pr}(\omega_k) = \frac{N_k}{N}, \tag{13}$$

$$\hat{\boldsymbol{\mu}}_k = [\hat{\mu}_{k,1}, \cdots, \hat{\mu}_{k,d}]^T,\ \hat{\mu}_{k,j} = \sum_{g_\ell = k} x_\ell / N_k, \ell = 1, \cdots, N_k \tag{14}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{N_k - 1} \sum_{g_\ell = k} (X_\ell - \hat{\mu}_k)(X_\ell - \hat{\mu}_k)^T \tag{15}$$

With the above estimations, we can assign a sample vector $\boldsymbol{x}$ to a class w_1 if $\delta_1(\boldsymbol{x}) > \delta_2(\boldsymbol{x})$. Then we can be in cases:

1. Assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} \Rightarrow$ LDA

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \frac{1}{2} \left[\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2^T \right] > \ln \frac{\Pr(\omega_1)}{\Pr(\omega_2)}$$

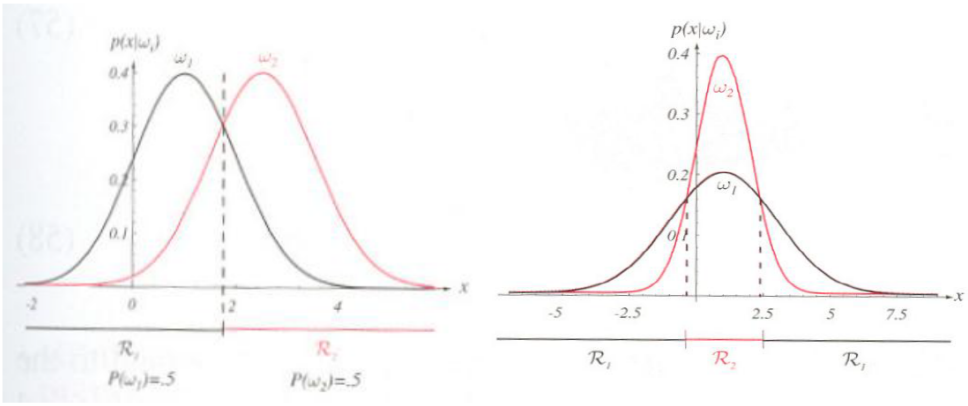
2. No assumption on the covariance matrix in different classes:

$$(x - \boldsymbol{\mu}_1) \boldsymbol{\Sigma}_1^{-1} (x - \boldsymbol{\mu}_1)^T - (x - \boldsymbol{\mu}_2) \boldsymbol{\Sigma}_2^{-1} (x - \boldsymbol{\mu}_2)^T + \frac{1}{2} \ln \frac{\boldsymbol{\Sigma}_1}{\boldsymbol{\Sigma}_2} > \ln \frac{\Pr(\omega_1)}{\Pr(\omega_2)}$$

Then substitute the estimates in (13) – (15), we can determine which class to be assigned. A compromise between LDA and QDA can be implemented using the regularized covariance matrix as

$$\boldsymbol{\Sigma}_k(\alpha) = \alpha \boldsymbol{\Sigma}_k + (1 - \alpha) \hat{\Sigma} \tag{16}$$

where α is a scalar in $[0, 1]$ and can be determined by cross validation.



$\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$
$\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2$

$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$
$\boldsymbol{\sigma}_1 \neq \boldsymbol{\sigma}_2$

Figure 1: Decision boundaries

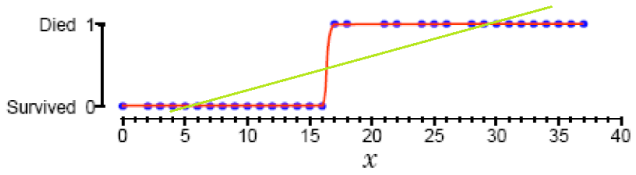


Figure 2: Logistic regression (blue curve) vs linear regression (green line)

3 Logistic regression

The motivation for logistic regression can be as follows:

1. Response is categorical,
2. The relation between response and predictors should be treated differently from general regression problem as shown in Figure 2.
3. The posterior probability of classes can be modeled via linear function of predictors.

For a multinormial (k) classification problem, we have

$$\begin{aligned}
\frac{\log \Pr(G=1 \mid X=x)}{\log \Pr(G=k \mid X=x)} &= \beta_{1,0} + \beta_1^T x \\
\frac{\log \Pr(G=2 \mid X=x)}{\log \Pr(G=k \mid X=x)} &= \beta_{2,0} + \beta_2^T x \\
&\vdots \\
\frac{\log \Pr(G=k-1 \mid X=x)}{\log \Pr(G=k \mid X=x)} &= \beta_{k-1,0} + \beta_{k-1}^T x
\end{aligned}$$

Then we can derive the close-forms as

$$\Pr(G=j \mid X=x) = \begin{cases} \frac{\exp(\beta_{j,0}+\beta_j^T x)}{1+\sum_{\ell=1}^{k-1}\exp(\beta_{\ell,0}+\beta_\ell^T x} & j=1,\cdots,k-1 \\ \frac{1}{1+\sum_{\ell=1}^{k-1}\exp(\beta_{\ell,0}+\beta_\ell^T x} & j=k \end{cases} \tag{17}$$

The estimation of the logistic regression parameters use the gradient based methods, e.g. Newton-Raphson method, which are not given here for simplicity.

4 Data description

A home equity loan is a type of loan in which the borrower uses the equity of his or her home as collateral. The description of the data is given in Figure 2. People with bad credit and without bad credit are separated into two groups. Then the Bayes Classifiers determine whether the customer paid and not paid home equity loans (BAD or not) using the features "DEROG","DELINQ", "CLAGE", "NINQ", "DEBTINC". The detailed explanation of the each variables can be found in `https://www.kaggle.com/ajay1735/hmeq-data`.

```
      BAD              LOAN          MORTDUE          VALUE          REASON       JOB           YOJ          DEROG          DELINQ
 Min.   :0.00000   Min.   : 1700   Min.   :  5076   Min.   : 21144   DebtCon:2181   Mgr    : 399   Min.   : 0.000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:11900   1st Qu.: 49267   1st Qu.: 71104   HomeImp: 886   Office : 545   1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.:0.0000
 Median :0.00000   Median :16800   Median : 67550   Median : 94267                  Other  :1173   Median : 7.000   Median :0.00000   Median :0.0000
 Mean   :0.09064   Mean   :18390   Mean   : 74405   Mean   :105171                  ProfExe: 833   Mean   : 9.062   Mean   :0.08934   Mean   :0.2214
 3rd Qu.:0.00000   3rd Qu.:23500   3rd Qu.: 92636   3rd Qu.:121258                  Sales  :  47   3rd Qu.:13.000   3rd Qu.:0.00000   3rd Qu.:0.0000
 Max.   :1.00000   Max.   :65800   Max.   :240782   Max.   :324987                  Self   :  70   Max.   :36.000   Max.   :6.00000   Max.   :8.0000
     CLAGE             NINQ            CLNO           DEBTINC
 Min.   :  8.055   Min.   : 0.0000   Min.   : 1.00   Min.   : 0.8381
 1st Qu.:118.462   1st Qu.: 0.0000   1st Qu.:16.00   1st Qu.:29.6590
 Median :174.475   Median : 0.0000   Median :21.00   Median :35.2377
 Mean   :177.545   Mean   : 0.9165   Mean   :21.88   Mean   :34.2138
 3rd Qu.:228.850   3rd Qu.: 1.0000   3rd Qu.:27.00   3rd Qu.:39.0576
 Max.   :440.421   Max.   :10.0000   Max.   :64.00   Max.   :78.6544
```

Figure 3: Summary of the home equity data

5 Analysis and results

The above classifiers are implemented in ".R". The comparison results is presented in Figure 3 in which the multiple logistic regression (MLR) performance best in terms of prediction error, the LDA and KNN comes second and third depends on the k values, The QDA and naive Bayes gives the worse prediction.

The reason for the worse performance of naive Bayes is its assumption that every feature is conditional independent which is not suitable for this data set, e.g. the " LOAN" strongly relates to the income "Value". The KNN has good performance when $k = 2$. The LDA give a universal variance to each class, which helps the classifier focus on the impact in different mean values. So LDA performs better than QDA. The logistic regression (MLR) performs best in this case, since the soft decision boundary that given by the logistic function.

We can analyze the data in following aspects:

1. comparing clients with who paid and not paid home equity loan as shown in Figure 4 which indicates the imbalance of the classes (the amount of"Bad" is much less).

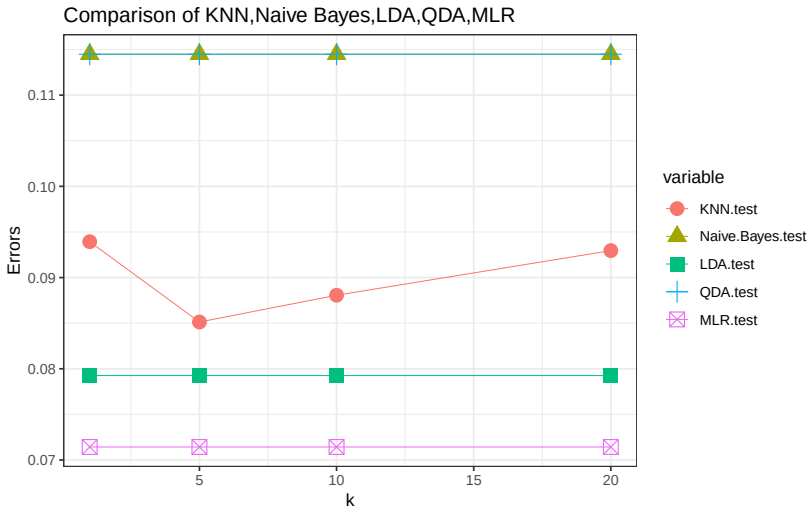


Figure 4: Comparison of KNN, NB, LDA, QDA, MLR

2. comparing the "Job" for each class as shown in Figure 5 and the ratio of "Bad" is in Figure 5 through which we know which job has the largest "BAD" rate.

The "Job" feature is a very important feature which is not consider in our case, thus resulted in the failure of Gaussian assumption for the features. To further improve the prediction accuracy, the segmentation by different "Jobs" is necessary. Then, for each group of people with the same job, the above classifier may works better.

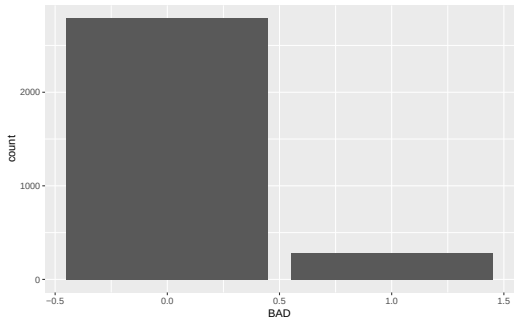


Figure 5: Bad or not

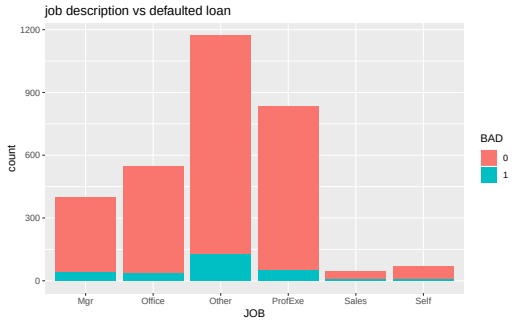


Figure 6: Number of "BAD" for each "Job"

Although the logistic regression showed a very good performance for the home equity loan data, we should pay attention when using it for the cases that the feature spaces are not linearly separable, e.g. the case presented in Figure 7.

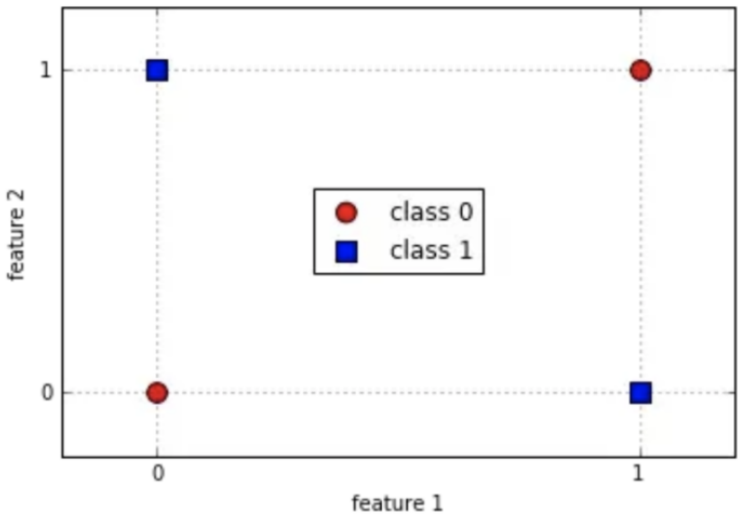


Figure 7: Linearly non-separable data