Data preparation: detect outliers using the Mahalanobis distance

Mingliang Wang

1 Mahalanobis distance

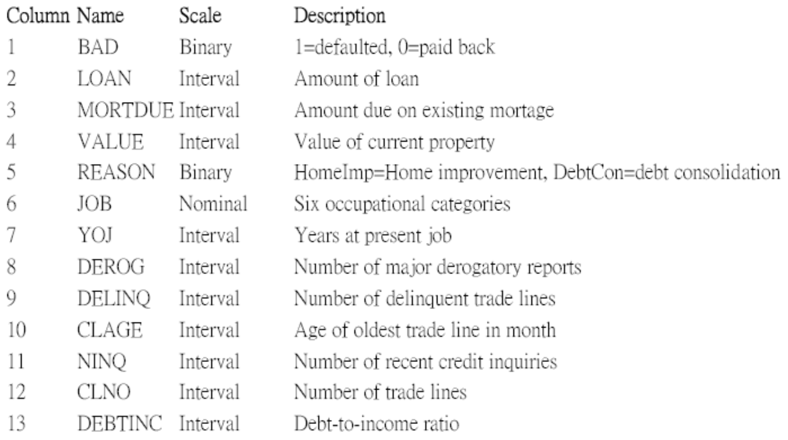
p-variate normal distribution, that is $X_1, \cdots, X_n$, i.i.d., with distribution $\mathcal{N}(\mu, \Sigma)$. According to distribution theory, the Mahalanobis distance

$$D^2 = (x - \bar{x})^T S^{-1} (x - \bar{x}) \sim \chi^2(p)$$

where x are one of the samples of $X_1, \cdots, X_n$ and $\bar{x}$ is the sample mean, S is the sample covariance. If the observation with D^2 greater than a pre-assigned level (say 99%) of a Chi-square distribution with p degrees of freedom, then this observation is an outlier.

2 Home equity loan data

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 1. People with bad credit and without bad credit are separated into two groups. In each group, the features are assumed to be multivariate distributed so that we can use the Mahalanobis distance to detect outliers. The implementation is given the R code “Outliers.R”. The result is presented in Figure 2.



Column	Name	Scale	Description
1	BAD	Binary	1=defaulted, 0=paid back
2	LOAN	Interval	Amount of loan
3	MORTDUE	Interval	Amount due on existing mortage
4	VALUE	Interval	Value of current property
5	REASON	Binary	HomeImp=Home improvement, DebtCon=debt consolidation
6	JOB	Nominal	Six occupational categories
7	YOJ	Interval	Years at present job
8	DEROG	Interval	Number of major derogatory reports
9	DELINQ	Interval	Number of delinquent trade lines
10	CLAGE	Interval	Age of oldest trade line in month
11	NINQ	Interval	Number of recent credit inquiries
12	CLNO	Interval	Number of trade lines
13	DEBTINC	Interval	Debt-to-income ratio

Figure 1: Description of home equity load data

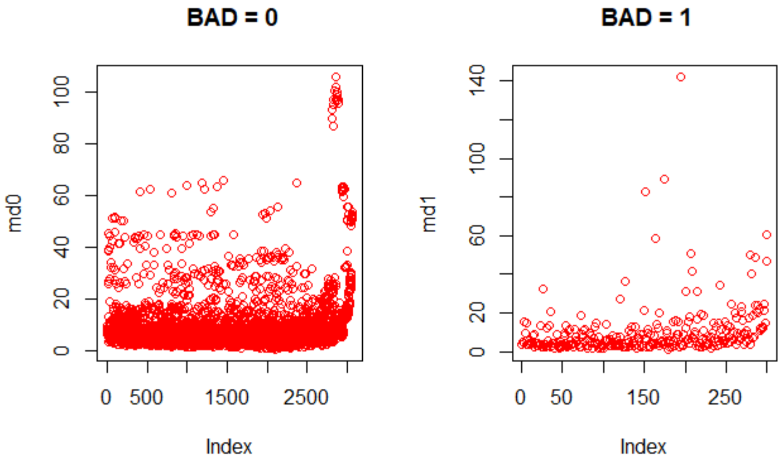


Figure 2: Mahalanobis distances for the two group, significance level = 99%, critical value = 23.21