Classification: Prediction protein localization sites

[redacted]

CE5310 Class Project

Introduction

Classification problem appeared in a board variety of real world problems. The classification result can be binary or multiclass. Binary classification problems have only two outcomes for each instance. Multiclass classification is to classify instances to three or more classes. For example, in medical studies, given the genomic data (DNA sequencing or DNA methylation levels) of one individual, researchers want to be able to tell whether the cancer patience of one of three cancer subtypes. In imaging recognition, given the image, we want the computers to put the image into correct category so the next steps can be done.

Dataset

The dataset for the multiclass classification problem in this project is obtained from UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Yeast). It is first created by

Kenta Nakai from Institue of Molecular and Cellular Biology, Osaka University, Japan. This

The eight explanatory variables plus the index variable are:

1. Sequence Name: Accession number for the SWISS-PROT database

2. mcg: McGeoch's method for signal sequence recognition.

3. gvh: von Heijne's method for signal sequence recognition.

4. alm: Score of the ALOM membrane spanning region prediction program.

5. mit: Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins.

6. erl: Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute.

7. pox: Peroxisomal targeting signal in the C-terminus.

8. vac: Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins.

9. nuc: Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

Methods

There are many existing methods for multiclass classifications. In the project, fuzzy rule based

Logistic model is widely used in modelling problems with a binary response variable. In the basic form, a logistic function is used. Like other forms of regression analysis, logistic regression makes use of one or more predictor variables that may be either continuous or categorical. Logistic regression is used for predicting dependent variables that take membership in one of a limited number of categories (treating the dependent variable in the binomial case as the outcome of a Bernoulli trial) rather than a continuous variable. To make the linear regression assumptions still hold, we need to apply a transformation on the response variable, so that it can be changed to a continuous one. To do that, binomial logistic regression calculates the odds of the event happening for different levels of each independent variable, and then takes its logarithm to create a continuous criterion as a transformed version of the dependent variable. The so-called logit transformation is defined as $\log\frac{\pi}{1-\pi}=\pi'$.

After the transformation, we estimate the model parameters using maximum likelihood estimation.

Ridge Regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When there are too many predictors, multicollinearity probably will be a problem. It is similar to the least squares estimation, where we want to minimize the residual adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. Ridge regression is the more popular compared to principle component regression.

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is in a standardized scale.

Ridge regression may be given a Bayesian interpretation. If we assume that each regression coefficient has expectation zero and variance 1/k, then ridge regression can be shown to be the Bayesian solution. Also, it can be shown that the ridge regression solution is achieved by adding rows of data to the original data matrix. These rows are constructed using 0 for the dependent variables and the square root of k or zero for the independent variables. One extra row is added for each independent variable. The idea that manufacturing data yields the ridge regression results has caused a lot of concern and has increased the controversy in its use and interpretation.

LASSO

The lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Like ridge regression, a penalty term is