Final Exam, Multivariate Analysis, Spring 2024

2024/6/12

Notes

1. You have three hours to complete this exam. Show all relevant work: partial credit will be given. Good luck!
2. Gather all your answers into an R Markdown (.Rmd) file. Then, upload this R Markdown file and its compiled html file to e3 (https://e3.nycu.edu.tw/) under the folder “final”.
3. If you have written part of your answers in the answer book, please also upload the scanned file of your written answers.
4. You can use “STAT-Wireless” to gain access to the internet.

Data

This exam will analyze the chest X-ray (CXR) images provided by the E-Da Hospital with disease labels for the classification model building. The images here have been pre-processed and summarized, resulting in 32 features (variables) to represent each image.

These chest X-ray images are either from disease-free subjects (normal), or may contain one of the two abnormalities in the chest.

To reduce the computational burden, we randomly selected 255 training observations and 57 testing observations for the following analysis. These two data sets (`CXR_train.csv` and `CXR_test.csv`) can be downloaded from e3 (https://e3.nycu.edu.tw/) under the folder “final”. They contain observations on

Column	Variable	Description
1	iid	Image unique ID
2	disease	Disease status of the image: 0=normal, 1=abnormality 1, 2=abnormality 2
3-34	F1-F32	32 image features

Questions

The objective is to design an abnormality detector that uses 32 features to predict image disease status. This is a supervised learning problem, with the outcome the class variable `disease` .

When performing supervised learning on high dimensional data, i.e., many predictors, the procedure is first to perform feature extraction to achieve dimension reduction and then apply machine learning

classification to predict data's class labels.

We thus propose the following four approaches to build the spam detector. **You are asked to implement these four approaches and evaluate their performance. Use `CXR_train.csv` to train the model and `CXR_test.csv` to obtain the misclassification rate.**

Approach 1

Feature extraction (10 points):

a. Pick up the features whose values are significantly different among different disease classes.

b. Since we need to perform the tests for multiple features simultaneously, the cut-off for the p-value, below which can be considered to be significant, needs to be set with care.

Machine learning classification:

a. (10 points) Use these selected features to run linear discriminant analysis and quadratic discriminant analysis.

b. (10 points) Use the 5-fold cross-validation (CV) to decide linear or quadratic discriminant analysis to be applied.

c. (10 points) Select linear or quadratic discriminant analysis to be used based on the CV misclassification rate.

Calculate the misclassification rate on test data (10 points)

Hints: To do the 5-fold cross-validation on `CXR_train.csv` , you need first to randomly split the data set into five subgroups (called folds).

```
CXR_train <- read.csv("CXR_train.csv", header=T)

set.seed(2024)
rows <- sample(nrow(CXR_train))
CXR_train_shuf <- CXR_train[rows,]
fold <- rep(1:5, rep(51, 5))
```

Then four folds are used to train the model and one fold to valid the performance of the model, with a total of five different train-valid combinations (five cross-validation rounds).

```
for (k in 1:5)
{
    train <- CXR_train_shuf[(fold!=k),]
    valid <- CXR_train_shuf[(fold==k),]

    # Train the model on train
    #   prediction on vaild
}

# Combine 5 prediction on valid
```

Approach 2

Feature extraction (10 points):

a. Perform PCA to obtain principal components (PCs), using the correlation matrix.

b. Select the number of PCs to explain at least 90% of the total variation.

Machine learning classification:

a. (10 points) Use these PCs to run the support vector machine classifier.

b. (10 points) Randomly split the data set `CXR_train.csv` in a 80:20 ratio for training and validation respectively.

c. (10 points) Use the training part to train the model with different kernel functions: linear, polynomial, radial, and sigmoid kernels, and calculate their misclassification rates on the validation part. Decide the “best” kernal function based on these misclassification rates.

Calculate the misclassification rate on test data (10 points)