MACQUARIE
University

Tutorial: PLS and PLS-DA

Benoit Liquet[*1]

[1]Macquarie University

[*]benoit.liquet-weiland@mq.edu.au

Contents

1 Prediction using PLS-DA

In this practice we will use the dataset `Sonar` from the `mlbench` R package. The Sonar data consist of 208 data points collected on 60 predictors. The goal is to predict the two classes M for metal cylinder or R for rock).

```
library(mlbench)
library(caret)
data(Sonar)
```

We first split the data into train/test data split

```
set.seed(107)
inTrain <- createDataPartition(
  y = Sonar$Class,
  ## the outcome data are needed
  p = .75,
  ## The percentage of data in the
  ## training set
  list = FALSE
)
```

By default, createDataPartition does a stratified random split of the data. To partition the data:

```
training <- Sonar[ inTrain,]
testing  <- Sonar[-inTrain,]
nrow(training)
[1] 157
nrow(testing)
[1] 51
```

(a) Here, a partial least squares discriminant analysis (PLSDA) model will be tuned over the number of PLS components that should be retained. Using a 10-fold cross-validation with 3 repetitions. Explore the argument `trainControl` form the `train()` function from caret package.

(b) Based on the previous results, decide the number of components to retain.

(c) Using your selected model, predict the label of the test data.

(d) Provide the confusion matrix

(e) Use the package `mixOmics` to perform the same analysis. You will use the function `plsda` from this package.

(f) Project the samples on the first two components. Use the function `plotIndiv()`

(g) Tune the number of component using a K-fold cross-validation approach by optimizing the area under the curve (auc). Help: use the `perf` function.

(h) Using your selected model, predict the label of the test data by using the `centroid.dist` distance. Provide the confusion matrix as well.

(i) Provide the roc curve evaluated on the test set using `auroc()` function.

2 PLS2 using mixOmics package

This data set contains the expression measure of 3116 genes and 10 clinical measurements for 64 subjects (rats) that were exposed to non-toxic, moderately toxic or severely toxic doses of acetaminophen in a controlled experiment.

```
library(mixOmics)
data(liver.toxicity)
X <- liver.toxicity$gene
Y <- liver.toxicity$clinic
help(liver.toxicity)
```

In this practice we will use PLS2 to model the relation between the X and Y variables. Here are the dimensions of the matrices that includes clinical parameters associated with liver failure.

```
dim(X)
[1]   64 3116
dim(Y)
[1] 64 10
```

(a) First start by tuning the number of components to select by using the `perf()` function and the Q^2 criterion using repeated cross-validation.

(b) Run the model with 2 components.

(c) The amount of explained variance can be extracted for each dimension and each data set:

(d) Using the `plotIndiv()` function, display the sample and metadata information using the arguments group (colour) and pch (symbol) to better understand the similarities between samples modelled with sPLS2. Interpret the results.

(e) Using the `perf()` function and a cross-validation approach provide the RMSE of the clinical variables.

(f) Provide the correlation circle plot by using a cut off of 0.5 to display high correlation.