Lab8: Mortality Prediction Challenge

Lydia Halter IDM 2022

Overview

Every day, doctors and nurses collect a lot of information about patients by asking questions and using adapted tools (stethoscope, syringe, sensors, etc.). This data is very useful to monitor health state, diagnose and choose treatments. It can also be used for statistical predictive analysis of how the patient will progress. Analysts at the *GetUHealthy Hospital* are currently using the Mean Method to predict mortality. The first part of this lab introduces you to the Mean Method using data from the Medical Information Mart for Intensive Care (MIMIC). The second part involves you in an online challenge to improve a system for predicting mortality of patients in the hospital's Intensive Care Unit (ICU).

Mortality Prediction Challenge

The main question of this challenge is: *How to predict the survival of a patient in the ICU given his or her medical record?* More specifically, you are asked to predict whether or not a patient will die during their stay at the hospital. Healthcare organizations use risk models one to help determine high risk patients that may need specialized care. Models like this are being used at Mount Sinai hospitals in New York City to predict if COVID-19 patients are likely to need ICU care and ventilation. They are also used to help determine which COVID-19 patients may safely recuperate at home. Medical experts use techniques like those in this lab, except they are based on more sophisticated deep learning models called *neural networks*.

Data

Data Description

The MIMIC training dataset contains information on 79,999 patients, represented by *categorical*, *binary* and *numerical* features (also called "variables"). Those features are for instance age, gender, ethnicity, marital status, as well as medical data such as blood pressure rate or glucose rate. There are a total of 342 variables. For the code solutions for Mean Method and LDA here, we just use the 49 numeric features. But feel free to use more of the variables in your creative analysis.

The class (or label) to be predicted is a binary variable telling if the patient died or not during while in the hospital. Fortunately for the patients, most of them don't die. Unfortunately for *us*, that leads to a class imbalance.

Data Preparation

This section reads in the training data and creates the "internal" training and validation set. A validation set acts like a test set. It has known class labels, and we use it to determine how good our classifier are. We call it a validation set to distinguish it from the "external" test set. The big difference is that you don't know the labels of this external test set. *Your job is to predict the labels of the external test set to enter the challenge.*

Your first step is to prepare the data. For the basic entry provided here, we only use numeric features that contain age and various biochemical indicators. *We do not scale.* For your entry, you can decide to use more features or scale. Just remember that you may need to convert non-numeric features to numbers first. Read the comments and the contest documentation to understand the files that are provided.

```
# Read in mortality data
mortdata <- read.csv("~/MATP-4400/data/mimic_synthetic_train.data", sep=' ', header=F)[ ,-1] # exclude

# read in the labels of the training data
labels <- factor(read.csv("~/MATP-4400/data/mimic_synthetic_train.solution", header=F)[ ,1], levels=c(0

#read in the feature types
feat.types <- read.csv("~/MATP-4400/data/mimic_synthetic_feat.type", stringsAsFactors = F, header=F)[,1]
# read in the feature names
cnames <- read.csv("~/MATP-4400/data/mimic_synthetic_feat.name", stringsAsFactors = F, header=F)[,1]

#Figure out which features are numerical
bool.feats <- feat.types=='Numerical'

mortdata <- mortdata[ ,bool.feats][ ,3:51] # only keep age and biochemical indicators
colnames(mortdata) <- cnames[bool.feats][3:51] # name the columns
```

The data was randomly divided into two sets consisting of 90% training and 10% validation. We'll use the validation set to estimate how well the classifier may work on future data.

```
# Split the data into training and testing sets
# train.size will be the number of data in the training set
n <- nrow(mortdata)
train.size <- ceiling(n*0.9)

#Set random seed to ensure reproducibility
set.seed(300)
trainID <- sample(n,train.size)

# For testing purposes...
head(trainID)
```

```
## [1] 66325 32700 59049 32897  4006  1312
```

```
#The training data is just the training rows
morttrain <- mortdata[trainID, ]

# Using -train gives us all rows except the training rows.
mortval <- mortdata[-trainID, ]
trainclass <- labels[trainID]
valclass <- labels[-trainID]

#We leave off the last column (the class label) to get the matrices of features
trainmatrix <- as.matrix(morttrain)
valtmatrix <- as.matrix(mortval)

# Let's look at the class labels of the training data
summary(trainclass)
```

```
## lived  died
## 69489  2511
```

```
summary(valclass)
```

```
## lived  died
##  7713   286
```

Exercise 1

Predict ICU Mortality using the Mean Method.

Construct a classifier on the training set using the Mean Method. This has been done for you.

```
Q <- trainmatrix
labels <- trainclass

# Make a vector y of class with 1 if M or -1 if B
y <- matrix(1,nrow=length(labels))
y[labels=='lived',1] <- -1

# Make Cplus the matrix containing the positive data
Cplus <- Q[y==1,]

# Make Cplus the matrix containing the positive data
Cneg=Q[y==-1,]

Mplus<-colMeans(Cplus)
Mneg<-colMeans(Cneg)
wMM<-(Mplus-Mneg)
wMM<-wMM/sqrt(sum(wMM^2)) # w/norm(w)

# Calculate threshold t
tMM <- ((Mplus+Mneg)%*%wMM)/2
tMM<-as.numeric(tMM) # make sure t is a scalar, not a matrix
```

Predict the training set using the Mean Method: Calculate the *Class 1* accuracy, the *Class -1* accuracy, and the *balanced* accuracy on the training set.

```
predict <- sign(Q%*%wMM-tMM)

confusion.matrix <- table(trainclass,predict)

kable(confusion.matrix, digits = 2)
```

	-1	1
lived	44568	24921
died	1153	1358

```
p_accuracy <- confusion.matrix[2,2]/(confusion.matrix[2,1]+confusion.matrix[2,2])
print(p_accuracy)
```

```
## [1] 0.5408204
```

```
n_accuracy <- confusion.matrix[1,1]/(confusion.matrix[1,1]+confusion.matrix[1,2])
print(n_accuracy)
```

```
## [1] 0.6413677
```

```
balanced <- (p_accuracy + n_accuracy)/2
print(balanced)
```

```
## [1] 0.591094
```

Predict the validationset using the Mean Method: Calculate the *Class 1* accuracy, the *Class -1* accuracy, and the *balanced* accuracy on the validation set.

```
Q <- valtmatrix
labels <- valclass

predict2 <- sign(Q%*%wMM-tMM)

confusion.matrix2 <- table(valclass,predict2)

kable(confusion.matrix2, digits = 2)
```

	-1	1
lived	4897	2816
died	117	169

```
p_accuracy2 <- confusion.matrix2[2,2]/(confusion.matrix2[2,1]+confusion.matrix2[2,2])
print(p_accuracy2)
```

```
## [1] 0.5909091
```

```
n_accuracy2 <- confusion.matrix2[1,1]/(confusion.matrix2[1,1]+confusion.matrix2[1,2])
print(n_accuracy2)
```

```
## [1] 0.6349021
```

```
balanced2 <- (p_accuracy2 + n_accuracy2)/2
print(balanced2)
```

```
## [1] 0.6129056
```

Exercise 2

Construct a classifier on the training set using the **LDA Method**.

```
train.df <- data.frame(trainmatrix, class=as.factor(trainclass))

lda.fit <- lda(class~., train.df,prior=c(1,1)/2)

w <- lda.fit$scaling

thresh <- ((lda.fit$means[1,] +lda.fit$means[2,])/2)%*%lda.fit$scaling
```

Predict the training set: Calculate the *Class 1* (Died) accuracy, the *Class -1* (Lived) accuracy, and the *balanced* accuracy on the training set.

```
predict3 <- predict(lda.fit,train.df)$class

confusion.matrix3 <- table(trainclass, predict3)

kable(confusion.matrix3, digits = 2)
```

	lived	died
lived	51584	17905
died	738	1773

```
p_accuracy3 <- confusion.matrix3[2,2]/(confusion.matrix3[2,1]+confusion.matrix3[2,2])
print(p_accuracy3)
```

```
## [1] 0.7060932
```

```
n_accuracy3 <- confusion.matrix3[1,1]/(confusion.matrix3[1,1]+confusion.matrix3[1,2])
print(n_accuracy3)
```

```
## [1] 0.7423333
```

```
balanced3 <- (p_accuracy3 + n_accuracy3)/2
print(balanced3)
```

```
## [1] 0.7242133
```

Predict the validation set using the LDA method: Calculate the *Class 1* (Died) accuracy, the *Class -1* (Lived) accuracy, and the *balanced* accuracy on the validation set.

```
val.df <- data.frame(valtmatrix, class=as.factor(valclass))

predict4 <- predict(lda.fit, val.df)$class
```

```
confusion.matrix4 <- table(valclass,predict4)

kable(confusion.matrix4, digits = 2)
```

	lived	died
lived	5754	1959
died	73	213

```
p_accuracy4 <- confusion.matrix4[2,2]/(confusion.matrix4[2,1]+confusion.matrix4[2,2])
print(p_accuracy4)
```

```
## [1] 0.7447552
```

```
n_accuracy4 <- confusion.matrix4[1,1]/(confusion.matrix4[1,1]+confusion.matrix4[1,2])
print(n_accuracy4)
```

```
## [1] 0.7460132
```

```
balanced4 <- (p_accuracy4 + n_accuracy4)/2
print(balanced4)
```

```
## [1] 0.7453842
```

Discuss how the LDA and Mean Method Classifiers are performing. Discuss the scores on the training and validation sets and compare. Which method is working better?

The LDA Method is performing significantly better than the Mean Method. With respect to the training set, Mean produces a balanced accuracy of 0.591 versus LDA's of 0.724. With respect to the validation set, Mean produces a balanced accuracy of 0.613 versus LDA's of 0.745.

Exercise 3

Make a model to predict mortality using R any way you like. *Use your creativity to make a better predictive model!* Can you use more data? Can you use different features? Can you use a different modeling algorithm? Would scaling help (don't forget to scale train, test, and validation datasets the same)?

You can get three points for making a different model. Earn up to 3 points extra credit for being creative and trying different things.

Predict the training set using your method. Calculate the *Class 1* (Died) accuracy, the *Class -1* (Lived) accuracy, and the *balanced* accuracy on the training set.

```
z.matrix <- scale(trainmatrix, center = TRUE, scale = TRUE)

z.df <- data.frame(z.matrix, class=as.factor(trainclass))

z.predict <- predict(lda.fit, z.df)$class
```

```
confusion.matrix5 <- table(trainclass,z.predict)

kable(confusion.matrix5, digits = 2)
```

	lived	died
lived	69489	0
died	2511	0

```
p_accuracy5 <- confusion.matrix5[2,2]/(confusion.matrix5[2,1]+confusion.matrix5[2,2])
print(p_accuracy5)
```

```
## [1] 0
```

```
n_accuracy5 <- confusion.matrix5[1,1]/(confusion.matrix5[1,1]+confusion.matrix5[1,2])
print(n_accuracy5)
```

```
## [1] 1
```

```
balanced5 <- (p_accuracy5 + n_accuracy5)/2
print(balanced5)
```

```
## [1] 0.5
```

Predict the validation set using the your method. Calculate the *Class 1* accuracy, the *Class -1* accuracy, and the *balanced* accuracy on the validation set.

```
z.matrix2 <- scale(valtmatrix, center = TRUE, scale = TRUE)

z.df2 <- data.frame(z.matrix2, class=as.factor(valclass))

z.predict2 <- predict(lda.fit, z.df2)$class

confusion.matrix6 <- table(valclass,z.predict2)

kable(confusion.matrix6, digits = 2)
```

	lived	died
lived	7713	0
died	286	0

```
p_accuracy6 <- confusion.matrix6[2,2]/(confusion.matrix6[2,1]+confusion.matrix6[2,2])
print(p_accuracy6)
```

```
## [1] 0
```

```
n_accuracy6 <- confusion.matrix6[1,1]/(confusion.matrix6[1,1]+confusion.matrix6[1,2])
print(n_accuracy6)
```

```
## [1] 1
```

```
balanced6 <- (p_accuracy6 + n_accuracy6)/2
print(balanced6)
```

```
## [1] 0.5
```

Exercise 4

- Compare performance of the Mean Method, LDA, and and your best shot model on the training and validation sets. Report how well the models do on the mortality data in terms of Class 1 accuracy, Class -1 accuracy, and balanced accuracy.

The "best shot" had a 0% accuracy regarding the positive (class 1) set, but a 100% accuracy regarding the negative (class -1) set, thus its balanced set falls in between the two at 50% (true for both training and validation sets). This is definitely not the ideal choice as it is sacrificing one side to benefit the other, whereas both of the other two methods each have a greater balanced accuracy (Mean at 0.591 and LDA at 0.724 for training; Mean at 0.613 and LDA at 0.745 for validation).

- Describe the predictive model you suggest for predicting ICU mortality on future patients. This could be any of the models that you have made in the previous sections. Explain why you selected this one.

I would suggest the LDA model as it has the highest balanced accuracy by a wide margin.

Exercise 5

For this project and the final project, you will being participating in an online challenge.

The Lab8 challenge is the "To Be or Not To Be Mortality Challenge" at https://codalab.lisn.upsaclay.fr/competitions/3073

The challenge asks you to predict survival of the patients in `~/MATP-4400/data/mimic_synthetic_test.data`

- The winner is the person who gets the highest score on this data.
- The label of this data are 0 for lived and 1 for died.
- Note this is a change from the -1, 1 format given above.
- The predictions must be in a file. The file **must** be called `mimic_synthetic_test.csv`.
- The submission is made by zipping this file, into a file which can be called by any name.

This exercise walks you through how to make an entry using the mean method. You should submit the mean method as your first entry. Everyones submissions will be the same, so you will know if you successfully submitted an entry. Then you can try entering again using your best shot method.

Done!

Prepare testing

First, we need to get the external testing data and prepare it just like we did the training data.

```r
# Read in the test data
mortdatatest <- read.csv("~/MATP-4400/data/mimic_synthetic_test.data", sep=' ', header=F)[ ,-1]

# Apply the same transformation to the test that we used on the training data.
mortdatatest <- as.matrix(mortdatatest[,bool.feats][ ,3:51]) # only keep age and numerical chemical dat
colnames(mortdatatest) <- cnames[bool.feats][3:51] # only keep age and numerical chemical data to avoid
```

Predict the external testing data

Now we predict the external testing data using the Mean method function made on the training data. (Note you may get a warning message here about object lengths but it is okay). We use the wMM and tMM to do the prediction.

```r
# This code is provided for you
# For the test set,
# Make a prediction using the mean method
ypredtestMM<-sign(mortdatatest%*%wMM-tMM)
# deal with 0 case
ypredtestMM[ypredtestMM==0]<- 0

# Switch from 1 (Died) and -1 (Lived) classes to the numbers 1 and 0 classes for the contest entry
ypredtestMM <- recode(as.factor(ypredtestMM), "1"=1, "-1" =0)

#verify all predictions are 0 or 1
summary(as.factor(ypredtestMM))
```

```
##     0     1
## 12583  7418
```

Prepare the testing data predictions for submission.

Now we create the entry file and zip it up for submission.
'BEWARE the output in ypredtest must be numbers with values equal to 0 or 1.

```r
# Here is the mean prediction file for submission to the website

write.table(ypredtestMM,file = "mimic_synthetic_test.csv", row.names=F, col.names=F)

#This automatically generates a compressed (zip) file
system("zip -u MMentry.csv.zip mimic_synthetic_test.csv")
```

Download the file `MMentry.csv.zip` to your laptop (using the RStudio **Export** command under the **More** menu item in **Files**) so you can enter the challenge.

Enter the challenge using the predictions in MMentry.csv.zip

Enter the Challenge and see your score.

- Go to https://codalab.lisn.upsaclay.fr/competitions/3073 and create an account using the **Sign Up** tab. Use your RCS ID as your username. Log in.

- Go to the **Participate** tab at: https://codalab.lisn.upsaclay.fr/competitions/3073
- Select **Submit/View Results**
- Describe your submission; make sure to mention your method
- Hit **Submit** to upload your file. Upload your `.zip` file from your computer.
- Hit **Refresh Status** to make sure that your submission was successful. *It can take a few minutes for the update to happen.*
- You know your entry was not in the proper format if you see `Failed` under status. You know that your entry worked if you see `Finished` under status.
- Look at the **View Scoring Output Log** tab to see how you did. You should see 60% balanced accuracy for the Mean Method.
- Go to the **Results** page to see how you did with respect to everyone.
- *If you see a negative number, then something went wrong and go back to the Output log to see what the problem was.*
- *Only your latest submission will count as your submission!*

Congratulations, you get three points if you successfully submit the Mean Method.

Write your score with the mean method here 0.6.

Exercise 6

Now you will create your own entry into the challenge. Surely you can do better than the Mean Method!

Note the Mean Method was made up for educational purposes. You should never use it real life! LDA and many other methods will probably always do better.

- Create a `.zip` file with the name of your choice containing `mimic_synthetic_test.csv` which contains your test data predictions.
- Use the same procedure for submitting as you did for the Mean Method. You should predict either `0` or `1` for each test data point.
- Export your entry file from Rstudio using the **Export** function under **More** in the **Files** pane.
- Create a `.zip` file and submit to the challenge website.

- If your score (left-hand column of results table) is a *negative* number, then you made some mistake.
  - Check out `View scoring output log` on the website to see what might have gone wrong.
  - Common mistakes are to not have the correct number of predictions — for example, if you predict the training set instead of the testing set — or to predict labels other than `0` and `1`.

NOTE: Your entry filename **must** be: `mimic_synthetic_test.csv` and it **must be zipped to submit.**

- Give your RCS ID/entry name and the score that you received on the website here:
- **RCS ID = _____ Balanced Accuracy = _____**
- Describe how you created your entry here.

The winning entry gets 5 points extra-credit.

Exercise 7

Machine learning classification models are being used to combat the COVID-19 pandemic. Find a published paper that does classification on some task related to COVID-19. (Google Scholar is a good place to look). Summarize the classification problem that the paper solves, the data set that is used, the machine learning method used, and the results. Write it up in a paragraph or two. Make sure to include the reference for the paper.

I used "COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images" located at https://www.nature.com/articles/s41598-021-88807-2.

The paper discusses the use of a "dimensionality reduction method" to enable a machine learning classifier to distinguish between a positive and negative COVID-19 case (this is because there are high similarities between covid and pneumonia regarding CXR imaging). The model produced results of 100% sensitivity and 96% precision through the use of a test set containing 84 CXR images