Homework

Course Name:

Statistics and Machine Learning

1. Using logistic regression model to classify Pima Diabetes data set. There are 9 variables and the last one --- “Diabetes”, coded “pos” and ”neg” is the label. You need to convert “pos” to 1 and “neg” to 0.

(1) Testing data set: Sampling 1/10 of negative examples and 1/10 from positive examples. Using it as your testing data set. (sampling without replacement)

(2) All the rest (9/10) data will be use for “training/modeling”. That is using these training data to fit a logistic model using function “glm” in R. Then use “predict” to predict the testing data. You should get predicted probabilities by using this function.

(3) Summarize you results as a statistical report; i.e. you need to explain you results. Please do not just report the accuracy. For example, you may say something about false positive/negative rate.

(4) Remember the accuracy depends on the your choice of cutting point. You should discuss how you choose it.

(5) Using R-package “ROCR” to draw the ROC curve for your predictions. (For both training and testing data sets; i.e. you will have a curve for training and a curve for testing.)

2. Using Fisher Linear Discriminant to do the same problem above; i.e. using linear discriminant analysis to construct classifier for Pima Diabetes Data. Answer the same questions above. Including There is no FDA package that you can load directly in R. Instead you need to install a package called “MASS”. There is a function called “lda” and you may use it for this problem. The details of usage, please refer to the document of MASS.

3. Now using linear regression to do the classification problem above. Now you will have continuous prediction values.

Compare the results obtained by using linear model with previous that of previous two methods from different viewpoints. For example, ROC curve, Accuracy, etc.

Homework

Hints and Suggestions:

General:

The ratio of the positive examples size to the negative example size of Pima Data set is not equal to 1. You may take this into consideration if you think it is necessary.

Although you are not biologists and bio-statisticians, please read the names of variables and try to use it interpret your results.

Variable selection is not required, but you can do it. I always think this is a part of data analysis/statistical modeling.

Please read the documents about R packages which you want to use and you are not limited to R or the packages I mentioned here. My codes may be buggy.

You are free to use any other software tools.

Some useful commands:

```
glm(diabetes~., data=your_training_data_set, family=binomial(link=logit))

table(true_label, predicted_label)

sample

index.pos <- which(pima[,9]=="pos") % will give you the index of pos examples.

pima[index.pos,]->positive_example % This will give you all the positive examples.
```

How to load data set in R:

Pima Diabetes data set can be download from UCI ML Bench and it is also a R package which has integrated many Machine Learning Bench Mark data set in it. You can use the following codes to install the data set into R in your computer:

```
> install.packages("mlbench")

> library(mlbench)

> data(PimaIndiansDiabetes)

> PimaIndiansDiabetes->pima  # Here you can use any name you want.

> pima
```

After these steps, pima data already load into your R workspace.

Examples of using ROCR

```
# plot a ROC curve for a single prediction run
# and color the curve according to cutoff.
data(ROCR.simple)
pred <- prediction(ROCR.simple$predictions, ROCR.simple$labels)
perf <- performance(pred,"tpr","fpr")
plot(perf,colorize=TRUE)
```