**STA6714-19Spring 0001-Homework for Chapter 12**

**Mithun Mohanraj**

**Identifying Good System Administrators. A management consultant is studying the roles played by experience and training in a system administrator's ability to complete a set of tasks in a specified amount of time. In particular, she is interested in discriminating between administrators who are able to complete given tasks within a specified time and those who are not. Data are collected on the performance of 75 randomly selected administrators. They are stored in the file SystemAdministrators.csv. Using these data, the consultant performs a discriminant analysis. The variable Experience measures months of full time system administrator experience, while Training measures number of relevant training credits. The dependent variable Completed is either Yes or No, according to whether or not the administrator completed the tasks.**

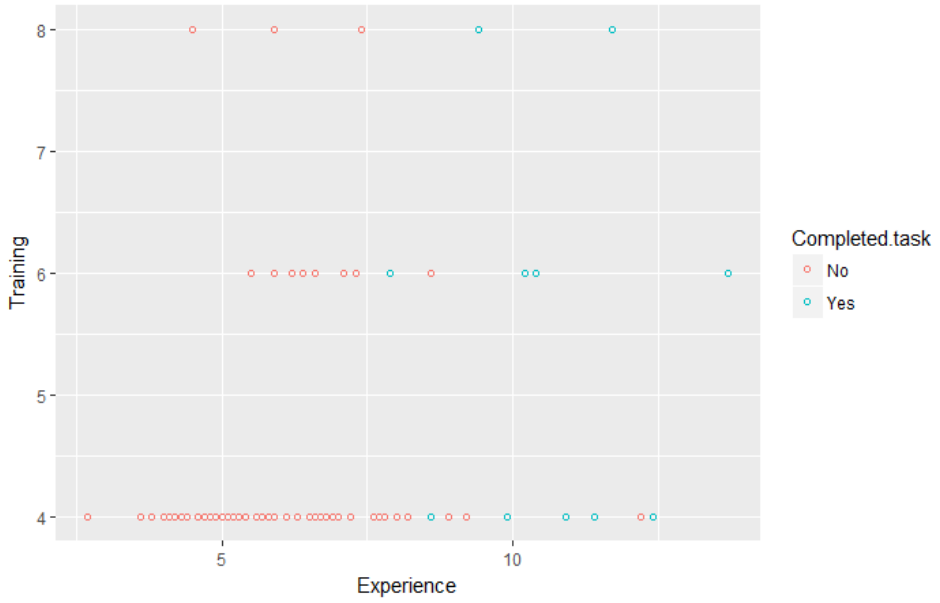
**A, Create a scatter plot of Experience vs. Training using color or symbol to differentiate administrators who completed the tasks from those who did not complete them. See if you can identify a line that separates the two classes with minimum misclassification.**

**R code :**

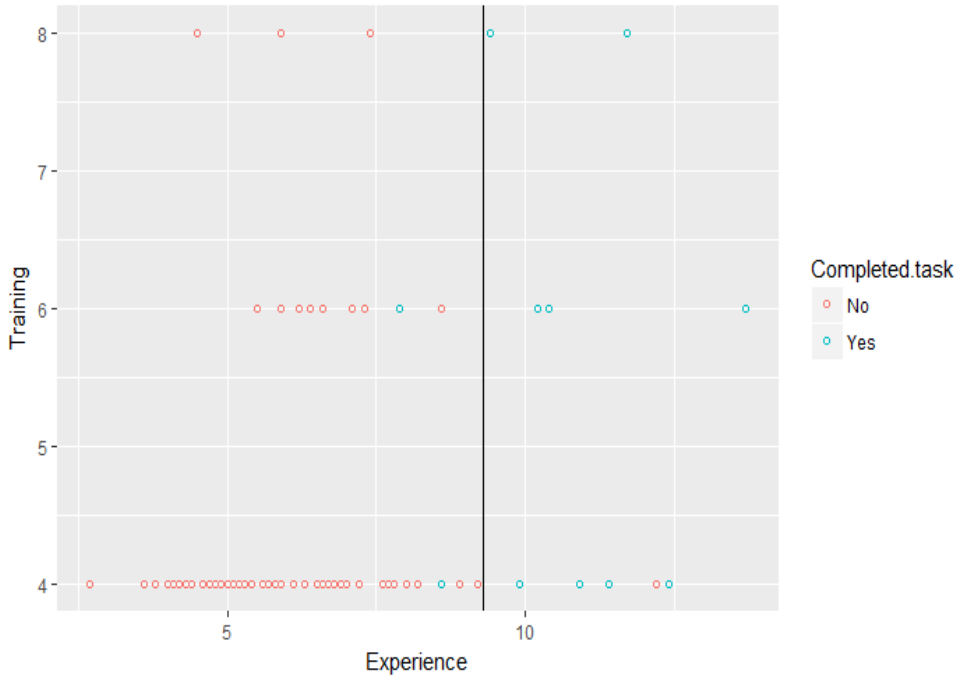
```
ggplot(dat, aes(x=Experience, y=Training, color=Completed.task)) +
geom_point(shape=1)
```

Scatter plot :

Experience vs training scatter plot to differentiate administrators for completed or not completed tasks using the colour .




**Line separating the two classes with minimum misclassifications is obtained at x-intercept = 9.3**




**This line separates the two classes with three misclassifications.**

**B,Run a discriminant analysis with both predictors using the entire dataset as training data. Among those who completed the tasks, what is the percentage of administrators who are classified incorrectly as failing to complete the tasks?**

**R code :**

```
da.reg <- linDA(dat[,1:2], dat[,3])

da.reg$functions

confusionMatrix(da.reg$classification, dat$Completed.task)
```

the constants and the classification functions for the two-class response variable are obtained as :

```
                    No         Yes
constant   -12.670275 -26.275402
Experience   1.952672   3.397371
Training     2.922710   3.066988
```

Confusion matrix for the entire data is obtained as :

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  58   5
       Yes  2  10
```

**Among 15 administrators , 5 of them were classified incorrectly as failing to complete the tasks .therefore, 33% of the administrators are classified incorrectly as failing to complete the tasks.**

**C,Compute the two classification scores for an administrator with 4 months of experience and 6 credits of training. Based on these, how would you classify this administrator?**

Based on constants and the classification functions , the classification scores are calculated as :

For not completed :

-12.670275 + (1.95 x 4) + (2.92 x 6)= 12.6497

For completed :

-26.275+ (3.397 x 4) + (3.066988 x 6) = 5.71488

**Since the classification score is high for "not completed" , the administrator will not complete the training with 4 months of experience and 6 credits of training.**

**D,How much experience must be accumulated by an administrator with 4 training credits before his or her estimated probability of completing the tasks exceeds 0.5?**

To get a propensity = 0.5 , the classification scores of both the classes have to be equal .

**When the administrator have 4 training credits , the administrator should have 9 months of experience to get a propensity of 0.5**

--12.670275 + (1.95 x 9) + (2.92 x 4) = 16.56

-26.275+ (3.397 x 9) + (3.066988 x 4) = 16.56

Propensity = exp(16.56)/ exp(16.56)+ exp(16.56)= 0.5

**Therefore , the administrator should have accumulated 9 months of experience with 4 training credits before his propensity of completing the task exceeds 0.5**

**E,Compare the classification accuracy of this model to that resulting from a logistic regression with cutoff 0.5.**

**Classification accuracy of LDA model is obtained from confusion matrix as :**

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  58   5
       Yes  2  10

               Accuracy : 0.9067
```

**R code for Logistic regression model :**

```
logit.reg <- glm(dat$Completed.task ~ ., data = dat, family = "binomial")

logit.reg.pred <- predict(logit.reg,dat, type = "response")

pred=ifelse(logit.reg.pred>0.5,"Yes","No")

conf=table(pred,dat$Completed.task)

confusionMatrix(conf)
```

**Classification accuracy of logistic regression model using 0.5 cutoff is obtained as:**

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  58   5
       Yes  2  10

               Accuracy : 0.9067
```

**Therefore, from the above results we can say that both LDA and logistic regression achieves the same accuracy of 90.67%**