DATA 621 Assignment 2

Kai Lukowiak

2018-03-14

Import Data

```
library(tidyverse)
library(knitr)
df <- read_csv('~/DATA621/Assignments/Assignment2/classification-output-data.csv')
sample_n(df, size = 5) %>% kable()
```

pregnant	glucose	diastolic	skinfold	insulin	bmi	pedigree	age	class	scored.class	scored.probability
0	111	65	0	0	24.6	0.660	31	0	0	0.1047958
2	84	50	23	76	30.4	0.968	21	0	0	0.0859556
6	80	66	30	0	26.2	0.313	41	0	0	0.1156596
0	124	70	20	0	27.4	0.254	36	1	0	0.1626446
2	90	70	17	0	27.3	0.085	22	0	0	0.0532099

Confusion Matrix

R's table function can be used to create a confusion matrix. For an more indepth explenation of this please see this excelent website.

```
x <- table(df$scored.class, df$class)
colnames(x) <- c('Actual Positive', 'Actual Negative')
rownames(x) <- c("Predicted Positive", 'Predicted Negative')
x %>% kable()
```

	Actual Positive	Actual Negative
Predicted Positive	119	30
Predicted Negative	5	27

The sum of the rows and columns can give insight into model performance. The rows represent the predicted values while the columns represent the actual values.

```
# Lable each row for easier computation:
a <- x[1, 1]; b <- x[1, 2]; c <- x[2, 1]; d <- x[2, 2]

Sensitivity <- a / (a + c)
Specificity <- d / (b + d)
PosPredVal <- a / (a + b)
NegPredVal <- d / (c + d)
Accuracy <- (a + d) / (a + b + c +d)
```

- The overall accuracy of the model (0.8066298) shows the total correct classification over all scores. This can be misleading because if 90% of the data is `positive`, a clasifier which only predicts `positive` will be 90% accurate.

- Sensitivity (0.9596774) of the model is the the ratio of predicted positives to total positives.
- Specificity (0.4736842) of the model is the accuracy of negative classification (the opposite of sensitivity).
- Positive and Negative predicted values (0.7986577 and 0.84375 respectively) are positive and negative values that were correctly specified.