Week10_1

Surenther

2024-10-30

Import CSV

```
# Import CSV
c_data <- read.table(file = "binary-classifier-data.csv", header = TRUE, sep = ",")
```

Logistic Regression

```
# Convert 'label' to a factor for binary logistic regression
c_data$label <- factor(c_data$label, levels = c(0, 1))

# Fit a binary logistic regression model
mymodel <- glm(label ~ x + y, data = c_data, family = binomial)
summary(mymodel)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = binomial, data = c_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.424809   0.117224   3.624  0.00029 ***
## x           -0.002571   0.001823  -1.411  0.15836
## y           -0.007956   0.001869  -4.257 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2075.8  on 1497  degrees of freedom
## Residual deviance: 2052.1  on 1495  degrees of freedom
## AIC: 2058.1
##
## Number of Fisher Scoring iterations: 4
```

Findings

_Intercept: The intercept's coefficient is 0.4248, which is statistically significant with a p-value of 0.00029 (indicated by "***"). This coefficient represents the log-odds of the outcome variable (label) being 1 when

both x and y are 0. A positive intercept suggests a baseline tendency towards label = 1._

Variable x: The coefficient for x is -0.0026, with a p-value of 0.1584. This p-value is above the common significance level (0.05), so we do not consider x to have a significant effect on the label outcome. This suggests that changes in x are not strongly related to changes in the probability of the outcome.

Variable y: The coefficient for y is -0.0080, which is statistically significant (p-value = 2.07e-05, indicated by "***"). This negative coefficient suggests that as y increases, the probability of label = 1 decreases, holding x constant.

In summary, the significant effect of y indicates it is an influential predictor for the outcome, while x does not significantly impact the outcome within this model.

Predict the accuracy

```
# Predict probabilities
predicted_probs <- predict(mymodel, type = "response")

# Convert probabilities to binary predictions with a 0.5 threshold
predicted_classes <- ifelse(predicted_probs > 0.5, 0, 1)

# Calculate accuracy
accuracy <- mean(predicted_classes == c_data$label)
accuracy
```

```
## [1] 0.4165554
```

Findings

An accuracy of 0.4166, or approximately 41.66%, means that the model correctly predicts the outcome about 42% of the time. This is relatively low, suggesting that the model may not be performing well in distinguishing between the two outcomes in the dataset. In general:

Low accuracy could indicate that the predictors (x and y in this case) may not be capturing enough information about the outcome variable. It might be beneficial to consider adding more relevant predictors, transforming existing variables, or trying alternative models to improve performance.