Training and Cross Validating a Logistic Regression Model

Tadros Salama

2/4/2021

Data containing results for US counties for the 2016 presidential election

`trump_win`: **Response variable** (1 = Trump won, 0 = Trump lost) `obama_pctvotes`: **Predictor variable** percent of votes cast for Obama in 2012

Randomly spliting data into a 70% training and 30% test set

```
set.seed(999)
n <- nrow(county_votes); n
```

```
## [1] 3112
```

```
round(0.7*n)
```

```
## [1] 2178
```

```
train_index <- sample(1:n, 2178)
county_votes_train <- county_votes[train_index, ]
county_votes_test <- county_votes[-train_index, ]
```

Fitting model using training data

```
glm1 <- glm(trump_win ~ obama_pctvotes, family = "binomial",
            data=county_votes_train)
summary(glm1)
```

```
##
## Call:
## glm(formula = trump_win ~ obama_pctvotes, family = "binomial",
##     data = county_votes_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3309   0.0030   0.0233   0.1231   2.1601
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    19.64711    1.20927   16.25   <2e-16 ***
## obama_pctvotes -0.36312    0.02289  -15.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1883.45  on 2177  degrees of freedom
## Residual deviance:  530.77  on 2176  degrees of freedom
```

```
## AIC: 534.77
##
## Number of Fisher Scoring iterations: 8
```

predictions for probabilities on test set

```
probs1 <- predict(glm1, newdata = county_votes_test, type = "response")
preds1 <- ifelse(probs1 > 0.5, 1, 0)
head(data.frame(probs1, preds1), n=15)
```

```
##         probs1 preds1
## 1  0.999954392      1
## 8  0.999433422      1
## 9  0.928154844      1
## 12 0.916188372      1
## 16 0.999975665      1
## 18 0.779429928      1
## 20 0.999995040      1
## 22 0.999999405      1
## 23 0.999909081      1
## 24 0.003424933      0
## 27 0.998055339      1
## 29 0.999988482      1
## 31 0.999997962      1
## 33 0.043763807      0
## 45 0.994136952      1
```

```
#confusion matrix
tb <- table(prediction = preds1,
            actual = county_votes_test$trump_win)
addmargins(tb)
```

```
##           actual
## prediction   0   1 Sum
##        0   125  22 147
##        1    24 763 787
##        Sum 149 785 934
```

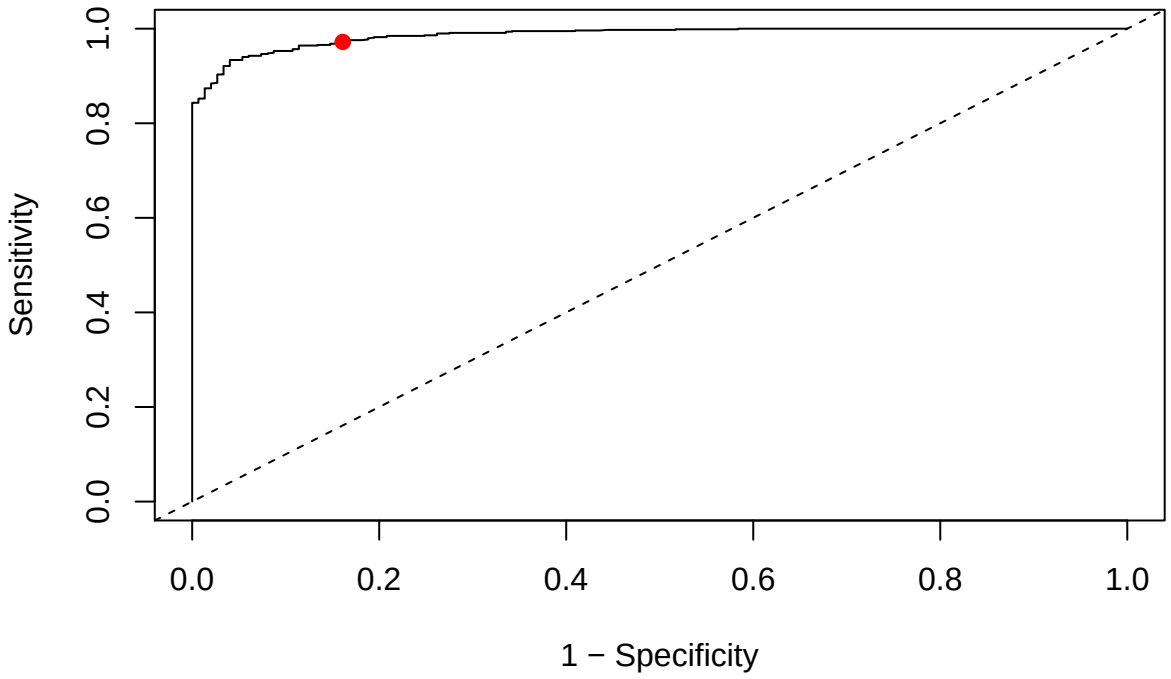
```
roc_obj <- roc(county_votes_test$trump_win, probs1)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
plot(1 - roc_obj$specificities, roc_obj$sensitivities, type="l",
     xlab = "1 - Specificity", ylab = "Sensitivity")

#red point corresponding to 0.5 threshold
points(x = 24/149, y = 763/785, col="red", pch=19)
abline(0, 1, lty=2) # 1-1 line
```




```
auc(roc_obj)
```

```
## Area under the curve: 0.9863
```