

Logistic Regression: Is this fish a Bream?

Gabe Wallon

2023-03-09

Affects of Group Prevalence on Logistic Model Accuracy

“When the prevalence within the binary response groups is asymmetric, we must be more careful with our classification threshold.” - my teacher

In this analysis I will predict whether a freshwater fish is a *Bream* based on the weight of the fish. The purpose of this analysis is to demonstrate how classification thresholds affect the accuracy of a model. Note that much of this code was taken from class examples and adapted to fit this context.

```
fish <- read.csv('/Users/gabrielwallon/Desktop/Data/fish_data.csv') %>%  
  mutate(Bream=ifelse(Species=='Bream',1,0))  
  
colnames(fish)
```

```
## [1] "Species" "Weight" "Length1" "Length2" "Length3" "Height" "Width"  
## [8] "Bream"
```

How many Bream are in this data set?

```
mean(fish$Bream)
```

```
## [1] 0.2464789
```

Therefore, if we were to guess that a given fish is not a Bream with no information about it at all, we would have around a 75% accuracy. Our model must have a greater accuracy than this.

Let's make a model.

```
model <- glm(Bream~Weight,data=fish,family='binomial')  
coefficients(summary(model))
```

```
##           Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -2.604943275 0.4046617509 -6.437335 1.215893e-10  
## Weight      0.003420773 0.0006872951  4.977154 6.452600e-07
```

Predicted probabilities of being a Bream:

```
fish_pred <- fish %>%  
  cbind(Bream_prob=predict(model,newdata=data.frame(Weight=fish$Weight),type='response'))
```

```
fish_pred %>%  
  summarize(Accuracy=mean((Bream_prob>0.5)==Bream))
```

```
## Accuracy  
## 1 0.7042254
```

With a classification threshold of 0.50, this model performs worse than if we were to guess without any information of the fish's weight. Thus we need to choose a new threshold.

```

#initiate vectors
thresholds <- seq(0,1,0.05)
accuracy <- rep(0,21)
j <- 1

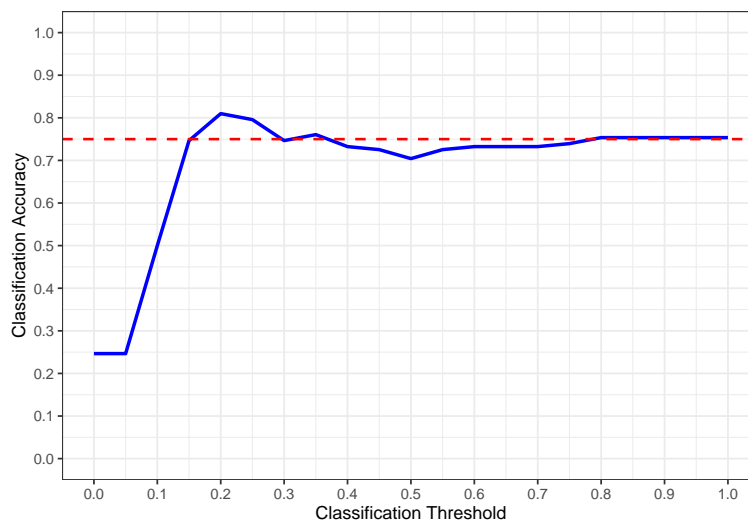
#loop through thresholds and calculate accuracy
for(i in thresholds){

  accuracy[j] <- fish_pred %>%
    summarize(mean((Bream_prob>i)==Bream)) %>%
    pull()

  j <- j+1
}

#graph accuracy versus threshold
data.frame(thresh=thresholds,accur=accuracy) %>%
  ggplot(aes(x=thresh,y=accur)) +
  geom_line(color='blue',size=1) +
  geom_hline(yintercept=0.75,linetype='dashed',
             color='red',size=0.75) +
  labs(x='Classification Threshold',
       y='Classification Accuracy') +
  scale_x_continuous(limits=c(0,1),breaks=seq(0,1,0.1)) +
  scale_y_continuous(limits=c(0,1),breaks=seq(0,1,0.1)) +
  theme_bw()

```



Can see that a classification threshold of around 0.2 allows the model to perform the best.

```

fish_class <- fish_pred %>%
  mutate(Bream_class=if_else(Bream_prob>0.5,1,0))

table(fish_class$Bream,fish_class$Bream_class)

```

```

##
##      0  1
##  0 93 14
##  1 28  7

```

Overall Accuracy = $100/142 = 0.704$ True Positive Rate = $7/35 = 0.20$ True Negative Rate = $93/107 = 0.869$

```
fish_class <- fish_pred %>%  
  mutate(Bream_class=if_else(Bream_prob>0.20,1,0))  
  
table(fish_class$Bream,fish_class$Bream_class)
```

```
##  
##      0  1  
##    0 84 23  
##    1  4 31
```

Overall Accuracy = $115/142 = 0.810$ True Positive Rate = $31/35 = 0.886$ True Negative Rate = $84/107 = 0.785$

By sacrificing a little of the true negative accuracy, our true positive accuracy increases dramatically after adjusting the classification threshold to 0.20. In the context of classifying a fish as a bream, we care most about overall accuracy because the consequences of a type 1 error, classifying a non-bream fish as a bream, and a type 2 error, failing to correctly classify a bream, are relatively equal.