Assignment 2 - 3373

Sebastian Doka and Justin Yee

2023-02-13

Question 1

Here are the functions for the error rate and the mean log loss

```
# inputs:
# ytrue is a vector of true binary responses (0 or 1)
# ypred is a vector of binary predictions (0 or 1)
# yprob is a vector of probabilistic predictions (between 0 and 1)

error.rate = function(ypred, ytrue){

  correct = 0

  for (i in 1:length(ypred)){
    if (ypred[i] == ytrue[i]){
      correct = correct + 1
    }
  }

  incorrect = length(ytrue)-correct
  err.rate = incorrect/length(ypred)

  return(err.rate)
}

mean.log.loss = function(yprob, ytrue){

  loss.vector = c()

  for (i in 1:length(yprob)){
    log.loss = -(ytrue[i])*log(yprob[i]) - (1 - ytrue[i])*log(1 - yprob[i]) #cross entropy loss functio
    loss.vector = c(loss.vector, log.loss)
  }

  mean.log.loss = mean(loss.vector)

  return(mean.log.loss)
}

# Test:
ytrue = c(0,1)
ypred = c(1,1)
yprob = c(0.8,0.55)


print(paste('Error rate = ', error.rate(ypred, ytrue)))
```

```
## [1] "Error rate =  0.5"
```

```
print(paste('Mean log loss = ', round(mean.log.loss(yprob, ytrue),3)))
```

```
## [1] "Mean log loss =  1.104"
```

Question 2

Load penguins dataset

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.4     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

A)

```
##
## Call:  glm(formula = sex ~ ., family = binomial, data = df)
##
## Coefficients:
##       (Intercept)   speciesChinstrap      speciesGentoo     bill_length_mm
##        -59.509989          -7.201288         -16.051746           0.626355
## flipper_length_mm        body_mass_g
##          0.056334           0.006712
##
## Degrees of Freedom: 332 Total (i.e. Null);  327 Residual
## Null Deviance:       461.6
## Residual Deviance: 161.9     AIC: 173.9
```

```
The error rate is 0.1081081.
The mean log loss is 0.2430287.
```

B)

The regression coefficient associated with flipper length is B4 = 0.0563. This means that for every one mm increase in flipper_length, the odds of having a male sex penguin are multiplied by exp(0.0563) = 1.057915.

C)

There are two regression variables associated with the species variable one for speciesChinstrap and one for speciesGentoo. For speciesChinstrap the regression coefficient is B1 = -7.2013 and for speciesGentoo the regression coefficient is B2 = -16.0517. From that we can tell that for 1 penguin in speciesChinstrap we have a decrease of exp(B1)= 7.4561588×10^{-4} in the sex of penguins. From that we can also tell that for 1 penguin in speciesGentoo we have a decrease of exp(B2)= 1.0686494×10^{-7} in the sex of penguins. But we also know that penguins is a binary variable with 0 as females and 1 as males. Therefore because both species have a negative coefficients the number of males decreases and the number of females increase for each species.As we can see that the speciesChinstrap has a lower regression coefficient value in this meaning more closer to 0 and to postive values therefore it has a higher proportion of male penguins than speciesGentoo.

D)

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

After splitting the dataset into two parts: train and test sets. We get that the error rate on the test set is 0.1492537 and the mean log loss on the test is 0.3207759

E)

After using the already data set from the previous and using a value of k=3, we get an error rate of 0.2537313

Question 3

A)

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
```

```
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

```
## [1] "Adelie"    "Chinstrap" "Gentoo"
```

```
## # weights:  18 (10 variable)
## initial  value 218.623845
## iter  10 value 31.450874
## iter  20 value 4.954607
## final  value 4.954607
## stopped after 20 iterations
```

The training error rate for the model is 0.0050251.

B)

```
## # weights:  18 (10 variable)
## initial  value 365.837892
## iter  10 value 66.463117
## iter  20 value 12.514139
## iter  30 value 7.743531
## iter  40 value 5.662963
## iter  50 value 4.350499
## iter  60 value 4.145847
## iter  70 value 4.093510
## iter  80 value 3.984724
## iter  90 value 3.903126
## iter 100 value 3.831009
## final  value 3.831009
## stopped after 100 iterations
```

The corresponding z-scores for each of the species are: SpeciesChinstrap: -24.8754276, speciesGentoo: -25.1595649 and for speciesAdelie: 0. The speciesAdelie is the reference level. The p values for each of the species are: SpeciesChinstrap: $1.5730374 \times 10^{-11}$, speciesGentoo: $1.1839675 \times 10^{-11}$ and for speciesAdelie: 1.

Question 4

Binomial logistic regression with weights $\mathbf{w} = \mathbf{w_1} - \mathbf{w_0}$:

$$P(Y=0) = \frac{\exp(-\mathbf{w} \cdot x)}{1 + \exp(-\mathbf{w} \cdot x)} \tag{1}$$

$$\tag{2}$$

$$P(Y=1) = 1 - P(Y=0) \tag{3}$$
$$= \frac{1}{1+\exp(-\mathbf{w}\cdot x)} \tag{4}$$

Now using multinomial logistic regression:

$$P(Y=0) = \frac{\exp(\mathbf{w_0}\cdot x)}{\exp(\mathbf{w_0}\cdot x)+\exp(\mathbf{w_1}\cdot x)} \tag{5}$$
$$= \frac{\exp((\mathbf{w_0}-\mathbf{w_1})\cdot x)}{\exp((\mathbf{w_0}-\mathbf{w_1})\cdot x)+1} \tag{6}$$
$$= \frac{\exp(-\mathbf{w}\cdot x)}{1+\exp(-\mathbf{w}\cdot x)} \tag{7}$$
$$\tag{8}$$

$$P(Y=1) = \frac{\exp(\mathbf{w_1}\cdot x)}{\exp(\mathbf{w_0}\cdot x)+\exp(\mathbf{w_1}\cdot x)} \tag{9}$$
$$= \frac{1}{\exp((\mathbf{w_0}-\mathbf{w_1})\cdot x)+1} \tag{10}$$
$$= \frac{1}{1+\exp(-\mathbf{w}\cdot x)} \tag{11}$$
$$\tag{12}$$

From this we can see that binomial logistic regression with weight vector $\mathbf{w}$ and multinomial logistic regression with two cases produce the same predictions.

Question 5

a) We have that $\sigma(z) = \frac{1}{1+\exp(-z)}$ as well as $(1-\sigma(z)) = \frac{\exp(-z)}{1+\exp(-z)}$. So when we take $\frac{d}{dz}\sigma(z) = \sigma\text{`}(z)$ we will get

$$\sigma\text{`}(z) = \frac{d}{dz}\left(\frac{1}{1+\exp(-z)}\right) \tag{13}$$
$$= -(1+\exp(-z))^{-2}(-\exp(-z)) \tag{14}$$
$$= \frac{-\exp(-z)}{-(1+\exp(-z))^2} \tag{15}$$
$$= \frac{\exp(-z)}{(1+\exp(-z))^2} \tag{16}$$
$$= \frac{1}{1+\exp(-z)} * \frac{\exp(-z)}{1+\exp(-z)} \tag{17}$$
$$= \sigma(z) * (1-\sigma(z)) \tag{18}$$

b) Following similar steps to part (a) we have that

$$
\begin{aligned}
\frac{\partial}{\partial \beta_j}\sigma(\beta_0+\beta_1 x_i^1+\cdots+\beta_p x_i^p) &= \frac{\partial}{\partial \beta_j}\left(\frac{1}{1+\exp(-\beta_0-\beta_1 x_i^1-\cdots-\beta_p x_i^p)}\right) && (19)\\
&= \frac{\partial}{\partial \beta_j}\left(\frac{1}{1+\exp(-\beta_j x_i^j)\exp(-\beta_0)\exp(-\beta_1 x_i^1)\ldots\exp(-\beta_p x_i^p)}\right) && (20)\\
&= -(1+\exp(-\beta_j x_i^j)(\ldots))^{-2} * (-x_i^j \exp(-\beta_j x_i^j)(\ldots)) && (21)\\
&= \frac{1}{1+\exp(-\beta_j x_i^j)(\ldots)} * \frac{x_i^j \exp(-\beta_j x_i^j)(\ldots)}{1+\exp(-\beta_j x_i^j)(\ldots)} && (22)\\
&= x_i^j * \frac{1}{1+\exp(-\beta_j x_i^j)(\ldots)} * \frac{\exp(-\beta_j x_i^j)(\ldots)}{1+\exp(-\beta_j x_i^j)(\ldots)} && (23)\\
&= x_i^j * p_i * (1-p_i) && (24)
\end{aligned}
$$

c) We have that $l_i = -y_i\log(p_i)-(1-y_i)\log(1-p_i)$, $p_i=\sigma(z_i)=\frac{1}{1+\exp(-z_i)}$, and $z_i=\beta_0+\beta_1 x_i^1+\cdots+\beta_p x_i^p$. We then take $\frac{\partial l_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j}(-y_i\log(p_i))+\frac{\partial}{\partial \beta_j}(-(1-y_i)\log(1-p_i))$ and solve each partial derivative separately.

$$
\begin{aligned}
\frac{\partial}{\partial \beta_j}(-y_i\log(p_i)) &= -y_i\frac{\partial}{\partial \beta_j}\log(p_i) && (25)\\
&= -y_i\left(\frac{\partial}{\partial \beta_j}(\log(1)-\log(1+\exp(-z_i)))\right) && (26)\\
&= -y_i\left(\frac{\partial}{\partial \beta_j}(-\log(1+\exp(-z_i)))\right) && (27)\\
&= y_i\left(\frac{1}{1+\exp(-z_i)}\right)\left(\frac{\partial}{\partial \beta_j}(1+\exp(-z_i))\right) && (28)\\
&= y_i\left(\frac{1}{1+\exp(-z_i)}\right)(-x_i^j\exp(-z_i)) && (29)\\
&= -x_i^j y_i(1-p_i) && (30)\\
&= x_i^j y_i(p_i-1) && (31)\\
& && (32)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \beta_j}(-(1-y_i)\log(1-p_i)) &= -(1-y_i)\left(\frac{\partial}{\partial \beta_j}\log(1-p_i)\right) && (33)\\
&= -(1-y_i)\left(\frac{\partial}{\partial \beta_j}\log(\exp(-z_i))-\frac{\partial}{\partial \beta_j}\log(1+\exp(-z_i))\right) && (34)\\
&= -(1-y_i)\left(\frac{1}{\exp(-z_i)}\frac{\partial}{\partial \beta_j}(\exp(-z_i))-\frac{1}{1+\exp(-z_i)}\frac{\partial}{\partial \beta_j}(1+\exp(-z_i))\right) && (35)\\
&= -(1-y_i)\left(\frac{-x_i^j\exp(-z_i)}{\exp(-z_i)}-\frac{-x_i^j\exp(-z_i)}{1+\exp(-z_i)}\right) && (36)\\
&= x_i^j(1-y_i)(1-(1-p_i)) && (37)\\
&= x_i^j(1-y_i)(p_i) && (38)
\end{aligned}
$$

Now we can add the results.

$$
\begin{aligned}
x_i^j y_i(p_i - 1) + x_i^j (1 - y_i)(p_i)) &= x_i^j (y_i p_i - y_i + p_i - y_i p_i) && (39) \\
&= x_i^j (p_i - y_i) && (40)
\end{aligned}
$$

Question 6

a) KNN classification would be good here as it would allow us to compare the taken image to known images of characters and evaluate how close (similar) they are. Considering most shipping labels are printed, this would be pretty dang accurate.

b) Multiple linear regression would be pretty good as we're trying to see which of our predictor variables (feature satisfaction) have the most impact on our response variable (overall satisfaction). So we'll be able to see which are significant and just how much each feature is weighted.

c) KNN classification would be good as it could find among the patient data a case very similar to the current patient and provide a starting point for medical staff.