Building A Model That Helps Locating Displaced People

Tom Lever

06/08/2023

In this project, we build a model that would help us locate people displaced by the earthquake in Haiti in 2010. More specifically, we build in a timely manner an accurate model that classifies pixels in geo-referenced aerial images of Haiti in 2010 as depicting vegetation, soil, rooftops, various non-tarp surfaces, and blue tarps. People whose homes had been destroyed by the earthquake often created temporary shelters using blue tarps. Blue tarps were good indicators of where displaced people lived.

Our training data was collected likely by applying a Region Of Interest (ROI) Tool to a high-resolution, orthorectified / geo-referenced image of Haiti in 2010. One ROI tool is described at https://www.l3harrisgeospatial.com/docs/regionofinteresttool.html. Classes may be assigned to pixels by defining Regions Of Interest.

Our training data frame consists of $63,241$ observations. Each observation consists of a class in the set $\{Vegetation,\ Soil,\ Rooftop,\ Various\ Non-Tarp,\ Blue\ Tarp\}$ and a pixel. A pixel is a colored dot. A pixel is represented by a tuple of intensities of color *Red*, *Green*, and *Blue* in the range 0 to 255.

According to https://www.esri.com/about/newsroom/insider/what-is-orthorectified-imagery/, an orthorectified image is an accurately georeferenced image that has been processed so that all pixels are in an accurate (x, y) position on the ground. "Orthorectified images have been processed to apply corrections for optical distortions from the sensor system, and apparent changes in the position of ground objects caused by the perspective of the sensor view angle and ground terrain."

We use 10-fold cross-validation to evaluate the performance of 5 classifiers. A classifier will classify a pixel as belonging to a class.

We load a data frame of classes and pixels based on an orthorectified image of Haiti at https://www.kaggle.com/datasets/billbasener/pixel-values-from-images-over-haiti?datasetId=1899167.

```
data_frame_of_classes_and_pixels <- read.csv(
    file = 'Data_Frame_Of_Classes_And_Pixels.csv'
)
head(x = data_frame_of_classes_and_pixels, n = 3)
```

```
#        Class Red Green Blue
# 1 Vegetation  64    67   50
# 2 Vegetation  64    67   50
# 3 Vegetation  64    66   49
```

TODO: Consider 5 number summaries and means and standard deviations of Red, Green, and Blue. and across classes.

According to https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/, "Below we use the `multinom` function from the `nnet` package to estimate a multinomial logistic regression model... we need to choose the level of our outcome that we wish to use as our baseline and specify this in the `relevel` function. Then, we run our model using `multinom`. The `multinom` package does not include `p`-value calculation for the regression coefficients, so we calculate `p`-values using Wald tests (here [two-tailed] z-tests)."

```
library(nnet)
factor_Class <- factor(x = data_frame_of_classes_and_pixels$Class)
data_frame_of_classes_and_pixels$Class <- relevel(
    x = factor_Class,
    ref = "Blue Tarp"
)
logistic_regression_model <- nnet::multinom(
    formula = Class ~ Red + Green + Blue,
    data = data_frame_of_classes_and_pixels
)
```

```
# # weights:  25 (16 variable)
# initial  value 101782.463020
# iter  10 value 77447.218043
# iter  20 value 29230.666252
# iter  30 value 24123.130528
# iter  40 value 23510.543863
# iter  50 value 23143.504515
# final  value 23072.994639
# converged
```

```
summary_of_logistic_regression_model <- summary(object = logistic_regression_model)
summary_of_logistic_regression_model
```

```
# Call:
# nnet::multinom(formula = Class ~ Red + Green + Blue, data = data_frame_of_classes_and_pixels)
#
# Coefficients:
#                  (Intercept)       Red      Green       Blue
# Rooftop            -3.001052 0.2394754 0.08565332 -0.3060410
# Soil              -11.994187 0.3168945 0.13645875 -0.4121712
# Various Non-Tarp   -2.917270 0.2516826 0.12434250 -0.3717489
# Vegetation         18.034812 0.1572812 0.38224636 -0.8164165
#
# Std. Errors:
#                  (Intercept)         Red      Green       Blue
# Rooftop          0.062157114 0.009033954 0.01213455 0.01209590
# Soil             0.086030376 0.008986432 0.01224512 0.01220922
# Various Non-Tarp 0.065428816 0.009095895 0.01226350 0.01221860
# Vegetation       0.003927454 0.011140036 0.01380552 0.01443139
#
# Residual Deviance: 46145.99
# AIC: 46177.99
```

```
coefficients <- summary_of_logistic_regression_model$coefficients
standard_errors <- summary_of_logistic_regression_model$standard.errors
z_scores <- coefficients / standard_errors
magnitudes_of_z_score <- abs(x = z_scores)
cumulative_density_function_values <- pnorm(q = magnitudes_of_z_score, mean = 0, sd = 1)
areas_in_one_tail <- 1 - cumulative_density_function_values
p_values <- areas_in_one_tail * 2
p_values
```

```
#                   (Intercept) Red        Green Blue
# Rooftop                     0   0 1.681544e-12    0
# Soil                        0   0 0.000000e+00    0
# Various Non-Tarp            0   0 0.000000e+00    0
# Vegetation                  0   0 0.000000e+00    0
```

The final negative log-likelihood l of our logistic regression model is $23,072.995$. The residual deviance $d = 2l = 46,145.990$.

TODO: Compare nested logistic regression models using residual deviance d.

The summary for our logistic regression model includes a data frame of coefficients. Each row of coefficients corresponds to a model equation. Interpreting the rows,

$$\begin{aligned} ln\left[\frac{P(Class=Rooftop)}{P(Class=Blue\ Tarp)}\right] &= \beta_{Rooftop,\ Intercept} + \beta_{Rooftop,\ Red}\ Red + \beta_{Rooftop,\ Green}\ Green + \beta_{Rooftop,\ Blue}\ Blue \\ &= -3.001 + 0.239\ Red + 0.0857\ Green - 0.306\ Blue \end{aligned}$$

$$\begin{aligned} ln\left[\frac{P(Class=Soil)}{P(Class=Blue\ Tarp)}\right] &= \beta_{Soil,\ Intercept} + \beta_{Soil,\ Red}\ Red + \beta_{Soil,\ Green}\ Green + \beta_{Soil,\ Blue}\ Blue \\ &= -11.994 + 0.317\ Red + 0.136\ Green - 0.412\ Blue \end{aligned}$$

Odds and relative risk are synonymous. An increase of 1 unit in predictor *Red* is associated with a change of 0.239 in the log odds of a pixel depicting a rooftop versus a pixel depicting a blue tarp. An increase of 1 unit in predictor *Blue* is associated with a change of -0.412 in the log odds of a pixel depicting soil versus a pixel depicting a blue tarp. Each predictor coefficient represents the log odds for a change of 1 unit in the predictor.

```
exp(x = coefficients)
```

```
#                    (Intercept)      Red    Green      Blue
# Rooftop          4.973474e-02 1.270582 1.089429 0.7363565
# Soil             6.180031e-06 1.372858 1.146208 0.6622109
# Various Non-Tarp 5.408115e-02 1.286188 1.132404 0.6895274
# Vegetation       6.798597e+07 1.170325 1.465573 0.4420128
```

The odds of a pixel depicting various non-tarp objects versus a pixel depicting a blue tarp is 1.132 for an increase of 1 unit in predictor *Green*. The odds of a pixel depicting vegetation versus a pixel depicting a blue tarp is 0.442 for an increase of 1 unit in predictor *Blue*.

The predicted probabilities for our first 3 observations and each class are presented below.

```
predicted_probabilities <- fitted(object = logistic_regression_model)
head(x = predicted_probabilities, n = 3)
```

```
#      Blue Tarp      Rooftop         Soil Various Non-Tarp Vegetation
# 1 2.521696e-06 3.992604e-05 1.049958e-07     4.741077e-05  0.9999100
# 2 2.521696e-06 3.992604e-05 1.049958e-07     4.741077e-05  0.9999100
# 3 1.633587e-06 3.224178e-05 8.961114e-08     3.933451e-05  0.9999267
```

The below data frame allows us to consider the changes in predicted probability associated with holding the intensity of color *Red* equal to the mean intensity of color *Red*, holding the intensity of color *Green* equal to the mean intensity of color *Green*, and increasing the intensity of color *Blue* from 0 to 255 inclusive linearly across 10 intensities. As the intensity of color *Blue* increases from 0 to 255, the predicted probability of a pixel depicting a blue tarp increases from 0 to 1 and the predicted probability of a pixel depicting vegetation decreases from 1 to 0.

```r
library(pracma)
mean_intensity_of_color_Red <- mean(data_frame_of_classes_and_pixels$Red)
mean_intensity_of_color_Green <- mean(data_frame_of_classes_and_pixels$Green)
linearly_spaced_intensities <- pracma::linspace(x1 = 0, x2 = 255, n = 10)
data_frame <- data.frame(
    Red = mean_intensity_of_color_Red,
    Green = mean_intensity_of_color_Green,
    Blue = linearly_spaced_intensities
)
predicted_probabilities <- predict(
    object = logistic_regression_model,
    newdata = data_frame,
    type = "probs"
)
predicted_probabilities
```

```
#       Blue Tarp      Rooftop         Soil Various Non-Tarp   Vegetation
# 1  3.355845e-45 7.744803e-24 7.139919e-19     2.351567e-20 1.000000e+00
# 2  3.730912e-35 1.476340e-17 6.728758e-14     6.966213e-15 1.000000e+00
# 3  4.147898e-25 2.814247e-11 6.341275e-09     2.063650e-09 1.000000e+00
# 4  4.605674e-15 5.357846e-05 5.968569e-04     6.105587e-04 9.987390e-01
# 5  1.504899e-07 3.001701e-01 1.653153e-01     5.315791e-01 2.935305e-03
# 6  2.239992e-03 7.660713e-01 2.085838e-02     2.108303e-01 3.929878e-09
# 7  9.423066e-01 5.525580e-02 7.437979e-05     2.363225e-03 1.487003e-16
# 8  9.999899e-01 1.005412e-05 6.690936e-10     6.682435e-08 1.419391e-26
# 9  1.000000e+00 1.723897e-09 5.671786e-15     1.780596e-12 1.276713e-36
# 10 1.000000e+00 2.955794e-13 4.807822e-20     4.744513e-17 1.148366e-46
```