Bayesian Multinomial Logistic Regression for Classifying Penguin Species from Penguin Features

by Zhihan Yang

STAT 340: Bayesian Statistics, Fall 2022, Carleton College

Abstract

This paper uses a non-informative Bayesian multnomial logistic regression analysis to study how well we can classify three penguin species based on just two biological features that can be reliably and safely measured from each individual penguin. We used palmerpenguins, a public available dataset, and the Gibb's sampler implemented in JAGS for MCMC computation. The resulting predictive model achieves over 97% classification accuracy on a held-out testing set, indicating the effectiveness of our approach.

1 Introduction

Wildlife identification, if done with high accuracy and minimal negative impact to the organism of interest, can be beneficial in many ways. Tools for wildlife identification can help researchers spot individuals of endangered species more efficiently, educate tourists about the species they encounter, help locals identify invasive species that are better off removed, etc. While there are powerful computer vision applications (some available through mobile phones) available for this purpose, data-driven computer vision algorithms often struggle in extreme situations (e.g., poor lighting and unusual weather conditions) that render the captured images "out-of-distribution" compared to training images. In this report, we take a different approach: we investigate how well we can classify three penguin species based on just two biological features that can be reliably and safely measured from each individual penguin. Our motivation is to provide a proof of concept and an example workflow for this approach that can be easily extended to provide identification for other plants and animals.

2 Data

We looked at the *penguins* dataset in the *palmerpenguins* package [1]. This data set contains the values of five biological features (bill length, bill depth, flipper length, body mass, and sex) for 344 penguins observed on the islands of Palmer Archipelago, Antarctica.

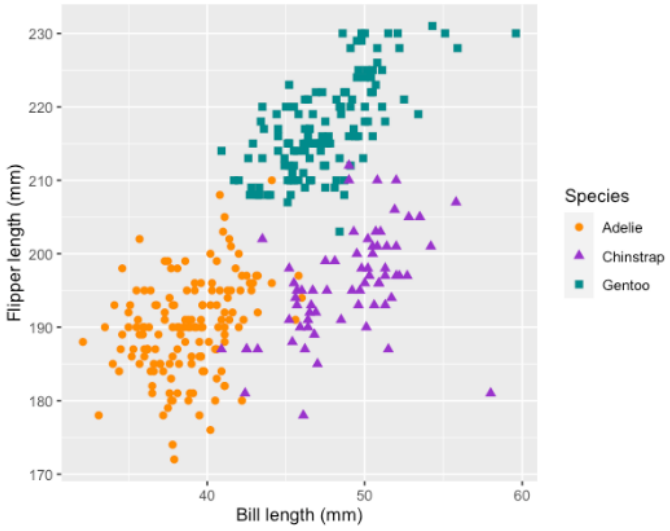

Figure 1. The three species of penguins in the data set appear to be linearly separable when viewed along the flipper length and the bill length dimensions.

For each penguin, the dataset additionally records its species (152 Adelie penguins, 68 Chinstrap penguins, and 124 Gentoo penguins), its location of observation (168 penguins observed on the Biscoe Island, 124 penguins observed on the Dream Island, and 52 observed on the Torgersen Island), and its year of observation (110 observed in 2007, 114 observed in 2008, and 120 observed in 2009). All these features were recorded from adult penguins at study nests found at the one-egg stage.

In this report, we are interested in predicting the *species* of a penguin given the values of its two biological features: *flipper length* and *bill length*. These two features were selected because the three species of penguins appear to be linearly separable when viewed along these two feature dimensions (Figure 1); linear separability is an important assumption for applying a multinomial logistic regression model because the decision boundary of this classification model is linear. Both features were collected using a ruler (with ± 1 mm in error) and have unit millimeter (mm). Out of the 344 penguins in the data set, 2 penguins have missing values for either species, flipper length, or bill length. These two penguins were removed prior to any analysis presented in this report.

3 Methods

3.1 Train-test split and feature standardization

Since we are interested in how well our model classifies unseen data, we split the

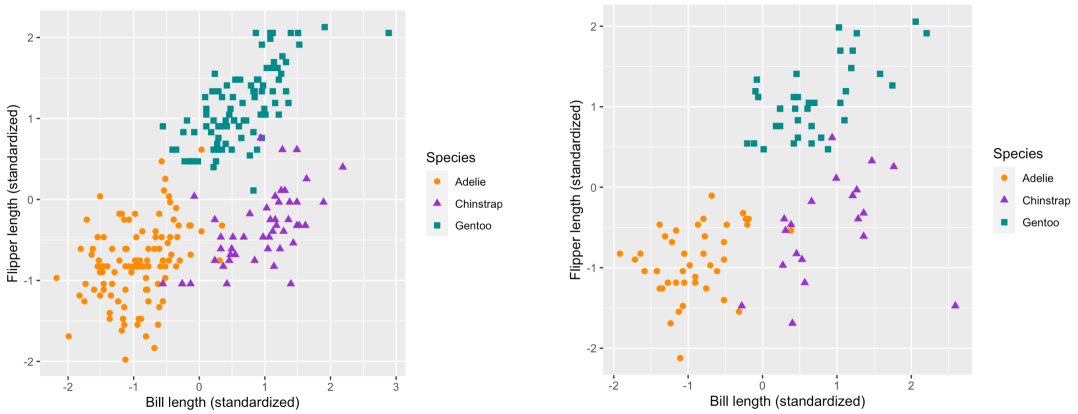

Figure 2. The standardized training set (left) and the standardized testing set (right).

data set into a training set and a testing set and only used the training set for Section 3.3; we discuss how posterior samples were used for classifying penguins in the testing set in Section 3.4. More specifically, we sampled 70% of examples from each class and these examples formed the training set; the remaining examples formed the testing set. We then standardized[1] the flipper lengths and bill lengths in the training set using their respective means and variances computed on the training set. We also standardized the flipper lengths and bill lengths in the testing set using their respective means and variances computed on the training (not testing!) set. The standardized training and testing sets are plotted in Figure 2. Notationally, we refer to the standardized training set as $\mathcal{D}_{\text{train}}$ and the standardized testing set as $\mathcal{D}_{\text{test}}$ in the mathematical equations in this report. For brevity, we will simply refer to the standardized training set as the training set and the standardized testing set as the testing set for the remainder of this report.

3.2 Model

Let y_i denote the species index of the i-th penguin, with 1 representing the Adelie species, 2 representing the Chinstrap species, and 3 representing the Gentoo species. Let $\vec{x}_i$ denote the feature (column) vector of the i-th penguin, with the first dimension being 1 (dummy value to multiple with the "intercept" terms), the second dimension being the penguin's standardized flipper length, and the third dimension being the penguin's stan-

1. Standardizing each predictor to have a mean of 0 and a standard deviation of 1 greatly improved the efficiency of the Gibbs sampler implemented in JAGS.

dardized bill length. Since the predictors are real-valued, the response y_i is categorical, and the three species appear to be linearly separable (Figure 1), we postulated that the multinomial logistic regression would be appropriate for modeling such data.

The multinomial logistic regression model has the following likelihood:

$$y_i | \, \vec{x}_i, \vec{\beta}_c, \vec{\beta}_g \overset{i.i.d.}{\sim} \text{Categorical}(\vec{\pi}_i),$$

where the entries of vector $\vec{\pi}_i$ are

$$\begin{aligned}
\pi_{i,1} &= \frac{\exp(0)}{\exp(0) + \exp(\vec{x}_i^T \vec{\beta}_c) + \exp(\vec{x}_i^T \vec{\beta}_g)} \\
\pi_{i,2} &= \frac{\exp(\vec{x}_i^T \vec{\beta}_c)}{\exp(0) + \exp(\vec{x}_i^T \vec{\beta}_c) + \exp(\vec{x}_i^T \vec{\beta}_g)} \\
\pi_{i,3} &= \frac{\exp(\vec{x}_i^T \vec{\beta}_g)}{\exp(0) + \exp(\vec{x}_i^T \vec{\beta}_c) + \exp(\vec{x}_i^T \vec{\beta}_g)},
\end{aligned}$$

and $\vec{\beta}_c = (\beta_{c,0}, \beta_{c,1}, \beta_{c,2})^T$ and $\vec{\beta}_g = (\beta_{g,0}, \beta_{g,1}, \beta_{g,2})^T$ are column vectors of parameters. Note that the numerator of $\pi_{i,1}$ is always 1 and does not depend any parameters; this is done to ensure the identifiability of $\vec{\beta}_c$ and $\vec{\beta}_g$. Finally, since we don't have a good prior understanding of the plausible values of these parameters, we place the following uninformative, independent (i.e., $\beta_{c,0} \perp \beta_{c,1} \perp \beta_{c,2} \perp \beta_{g,0} \perp \beta_{g,1} \perp \beta_{g,2}$) priors on them:

$$\begin{aligned}
\beta_{c,0} &\sim \mathcal{N}(0, 10) \\
\beta_{c,1} &\sim \mathcal{N}(0, 10) \\
\beta_{c,2} &\sim \mathcal{N}(0, 10) \\
\beta_{g,0} &\sim \mathcal{N}(0, 10) \\
\beta_{g,1} &\sim \mathcal{N}(0, 10) \\
\beta_{g,2} &\sim \mathcal{N}(0, 10),
\end{aligned}$$

where the second entry of each normal distribution is its standard deviation. Given the likelihood and priors, the unnormalized posterior distribution is obtained by Bayes' rule:

$$\begin{aligned}
& p\big(\vec{\beta}_c, \vec{\beta}_g | \, \mathcal{D}_{\text{train}}\big) \\
\propto \; & p\big(\mathcal{D}_{\text{train}} | \, \vec{\beta}_c, \vec{\beta}_g\big) \, p\big(\vec{\beta}_c, \vec{\beta}_g\big) \\
= \; & \left(\prod_{i=1}^{|\mathcal{D}_{\text{train}}|} p\big(y_i | \, x_i, \vec{\beta}_c, \vec{\beta}_g\big) \right) \left(\prod_{p=1}^{3} \mathcal{N}(\beta_{c,p} | \, 0, 10) \right) \left(\prod_{p=1}^{3} \mathcal{N}(\beta_{g,p} | \, 0, 10) \right).
\end{aligned}$$

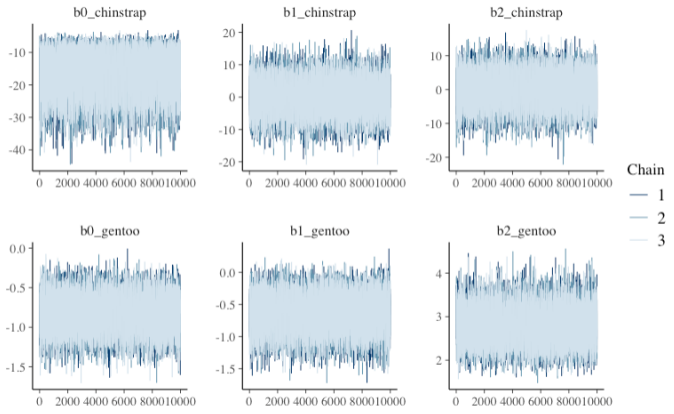

Figure 3. Trace plots for all parameters in the model. The notation here slightly differs from the notation used in Section 3.2. For example, here `b0_chinstrap` corresponds to $\beta_{c,0}$

3.3 Posterior sampling, convergence and posterior predictive check

Given the likelihood and priors defined in Section 3.2, there's no closed-form solution for the induced posterior distribution. Therefore, we used JAGS's R interface runjags, which internally runs a Gibbs sampler, to obtain posterior samples. Importantly, we only pass in the training set. We ran three independent chains, each with 10000 adaptation steps, 10000 warmup steps and 10000 sample steps with a thinning parameter of 5 (50000 steps were run for each chain after adaptation and warmup). Finally, we checked the trace plots for all parameters to verify convergence of the chains (Figure 3); all chains for all parameters appear to have reached the stationary distribution and show low autocorrelation.

To check whether the model is appropriate for the training set, we also performed a posterior predictive check by randomly reclassifying the predictor values using the posterior predictive distribution. To obtain one *reclassification* of the training set, we needed to obtain one random draw from

$$p(y_i|\,\vec{x}_i, \mathcal{D}_{\text{train}}) = \iint_{\Omega} p\big(y_i|\,\vec{x}_i, \vec{\beta}_c, \vec{\beta}_g\big) p\big(\vec{\beta}_c, \vec{\beta}_g|\,\mathcal{D}_{\text{train}}\big)\, d\vec{\beta}_c\, d\vec{\beta}_g$$

for each $\vec{x}_i$ in the training set; in practice, to bypass the intractable integral above, we sampled this draw from the categorical distribution parametrized[2] by a randomly selected[3]

2. Indirectly through class probabilities.

3. A posterior sample is randomly selected for *each* $\vec{x}_i$; we did not use the same posterior sample for all $\vec{x}_i$.

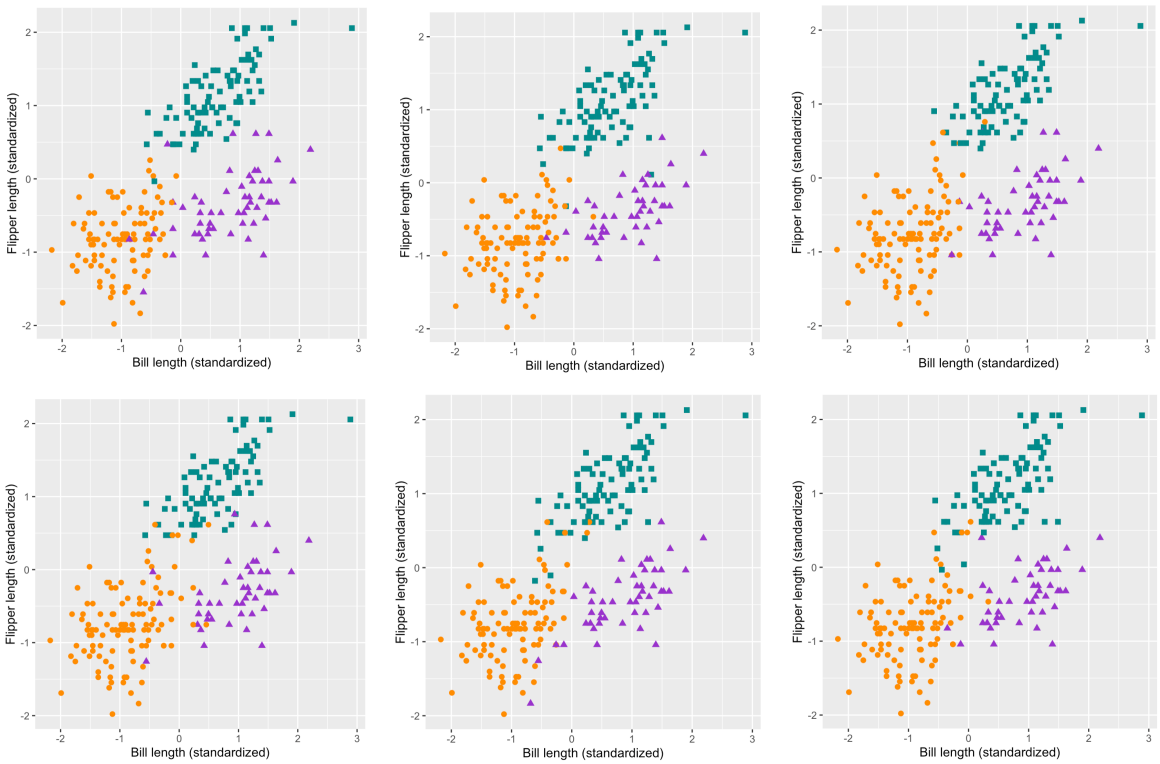

Figure 4. 6 reclassifications of the training set using the posterior predictive distribution. (Legend: orange for Adelie, green for Gentoo, purple for Chinstrap)

posterior sample $\left(\vec{\beta}_c^{(j)}, \vec{\beta}_g^{(j)}\right)$ and $\vec{x}_i$. Six reclassifications are shown in Figure 4. We see that the observed classification (left of Figure 2) could have been plausibly generated by this reclassification procedure, which means that the multinomial logistic regression model is appropriate for the training set.

3.4 Parameter inference and prediction

After ensuring model appropriateness, we then estimated the parameters of the model by constructing 95% credible intervals for each parameter using its 30000 posterior samples. To predict a penguin's species y given its feature vector $\vec{x}$, we theoretically need to compute (for each and every k):

$$\begin{aligned} p(y=k|\,\vec{x},\mathcal{D}_{\text{train}}) &= \iint_{\Omega} p(y=k|\,\vec{x},\vec{\beta}_c,\vec{\beta}_g)\,p(\vec{\beta}_c,\vec{\beta}_g|\,\mathcal{D}_{\text{train}})\,d\vec{\beta}_c\,d\vec{\beta}_g \\ &= \iint_{\Omega} \pi_k(\vec{x},\vec{\beta}_c,\vec{\beta}_g)\,p(\vec{\beta}_c,\vec{\beta}_g|\,\mathcal{D}_{\text{train}})\,d\vec{\beta}_c\,d\vec{\beta}_g, \end{aligned} \tag{1}$$

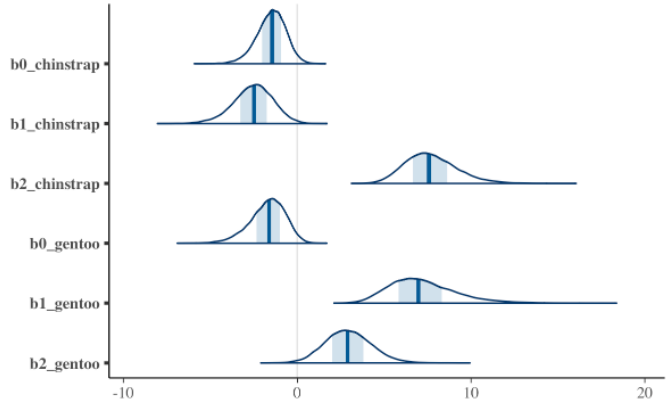

Figure 5. The empirical distribution of all 30000 posterior samples for each parameter. The dark vertical lines mark the medians; the light blue vertical bands mark the 50% credible intervals centered at the medians.

(where π_k is the probability of the k-th species as a function of $\vec{x}$, $\vec{\beta}_c$ and $\vec{\beta}_g$) and then perform $\arg\max_k p(y=k|\,\vec{x}, \mathcal{D}_{\text{train}})$ to select the species with the highest posterior probability as the classification result for $\vec{x}$. However, since the integral above is intractable, we resort to a Monte Carlo approximation of the integral using the 30000 posterior samples $\left\{\left(\vec{\beta}_c^{(j)}, \vec{\beta}_g^{\,(j)}\right)\right\}$ as follows:

$$p(y=k|\,\vec{x}, \mathcal{D}) \approx \frac{1}{30000} \sum_{j=1}^{30000} \pi_k\left(\vec{x}, \vec{\beta}_c^{(j)}, \vec{\beta}_g^{\,(j)}\right) \quad \forall k.$$

4 Results

Figure 5 shows the empirical distribution of posterior samples for each parameter. For the Chinstrap species, the middle 95% credible is $(-3.28, 0.0128)$ for the "intercept" term $\beta_{c,0}$, $(-5.00, -0.429)$ for the first "slope" term $\beta_{c,1}$, and $(5.16, 11.0)$ for the second "slope" term $\beta_{c,2}$. For the Gentoo species, the middle 95% credible is $(-4.01, 0.0267)$ for the "intercept" term $\beta_{g,0}$, $(4.14, 11.6)$ for the first "slope" term $\beta_{g,1}$, and $(0.460, 5.66)$ for the "second" slope term $\beta_{g,2}$. Since there are more than two classes, interpreting these parameters is significantly more difficult than a binary logistic regression model, so we leave this for future work. Figure 6 visualizes $p(y=1|\,\vec{x}, \mathcal{D}_{\text{train}})$ evaluated on a fine grid of values using Monte Carlo approximation of Equation 3.4; recall that index 1 corresponds to

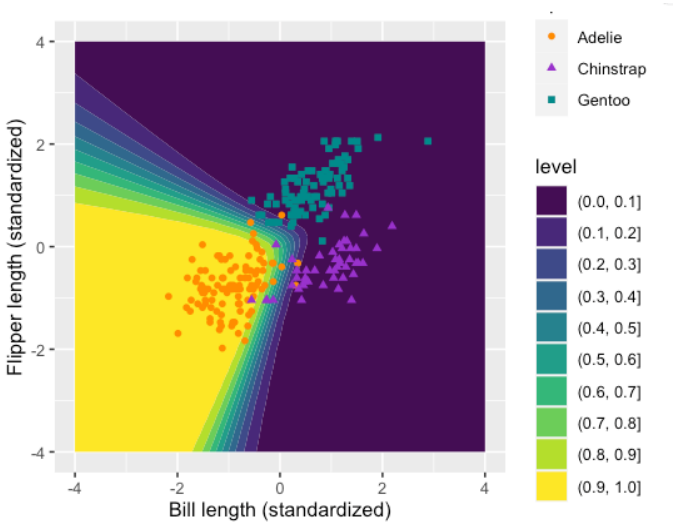

Figure 6. $p(y=1|\,\vec{x}, \mathcal{D}_{\text{train}})$ evaluated on a fine grid of points using Monte Carlo approximation of Equation 3.4. The index 1 corresponds to Adelie. The data plotted is the training set. We see that, as expected, examples in the Adelie species are in the region with higher $p(y=1|\,\vec{x}, \mathcal{D}_{\text{train}})$.

Adelie. We see that, as expected, training examples in the Adelie species are in the region with higher $p(y=1|\,\vec{x}, \mathcal{D}_{\text{train}})$. Finally, selecting the species with the highest posterior predictive probability leads to a 97.03% classification accuracy on the testing set.

5 Discussion and conclusion

This report aims to investigate how well we can classify three penguin species based on just two biological features that can be reliably and safely measured from each individual penguin. In particular, we performed a Bayesian multinomial logistic regression analysis on the training set and evaluated the resulting predictive model on a held-out testing set. The high classification accuracy on the testing set indicates that our approach has been highly successful. The presenteed workflow can be easily extended to another data set.

There are a few limitations that are worthy of further study. To begin, the decision boundary of the multinomial logistic regression model is linear, meaning that it is only appropriate for data where the classes are linearly separable with respect to the features. Future work could look at using more sophisticated models (e.g., one with polynomials features) that allow for more flexible decision boundaries. However, increasing the number of parameters usually come with a higher computational cost in posterior sampling and

a worse interpretability of the parameters in the model; interpreting parameters in the model presented is already difficult. Future work could also look at incorporating more prior knowledge into the model.

References

[1] Horst, A. M., Hill, A. P., & Gorman, K. B. (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. doi:10.5281/zenodo.3960218