STAT 535: Chapter 14: Naive Bayes Classification

David B. Hitchcock
E-Mail: hitchcock@stat.sc.edu

Spring 2022

Goal of Classification

- The basic goal of **classification** in statistics is to predict the category of a categorical response variable Y, based on values of one or more predictor variables $(X_1, X_2, \ldots, X_p)$.
- We did this already in logistic regression, when we used our model to classify new observation(s) as having $Y = 1$ or $Y = 0$, based on their predictor value(s).
- However, in logistic regression, our categorical response variable can have only two categories.
- In some data sets, a categorical response variable could have several (more than two) classes.

Example of a Multicategory Response Y

- Consider three types of Antarctic penguins (Adelie, Chinstrap, and Gentoo).
- Our goal is to classify an observed penguin into one of these three categories, based on measurements on three predictor variables: X_1 = weight (= 1 if above average, = 0 if below average), X_2 = bill length (in mm), and X_3 = flipper length (in mm).
- The `penguins_bayes` data frame contains measurements on 344 Antarctic penguins for whom the species is known.
- In this sample, there are 152 Adelies, 68 Chinstraps, and 124 Gentoos.

Possible Prior Specifications

- One approach is to assume the proportions of each type in the sample reflect the proportions in the general population (this is maybe the most common approach).
- Another approach is to specify subjective prior probabilities of a new observation belonging to each class.
- A noninformative approach would assign equal prior probabilities for each class, but this may not be the best idea unless the category proportions were actually relatively similar in the population.

Naive Bayes Classification Compared to Logistic Regression

- We have seen how logistic regression can be used as a classifier when the response variable is binary.
- **Naive Bayes Classification** is another classification method that has certain advantages:
- It can classify categorical response variables Y with two or more categories
- Doesn't require much theory beyond Bayes' Rule
- Computationally efficient, not requiring MCMC simulation

Example of Naive Bayes Classification with One Categorical Predictor

- Consider using only the categorical predictor "above_average_weight" to classify a new penguin into one of the three species.
- Note two penguins have missing predictor values, so we use only 342 penguins in the remainder of the analysis.
- A bar plot shows the most likely species to be below average weight ($X_1 = 0$) is Chinstrap.
- Is we encounter a penguin that is below average weight, should we classify it as Chinstrap?
- Be careful: Chinstrap is the rarest type of penguin to encounter in general.

Bayes' Rule for Classification with One Categorical Predictor

- Recall Bayes' Rule: The probability that a categorical response takes value y^*, given a particular value of categorical predictor X_1, is:

$$p(y^* \mid x_1) = \frac{\text{prior} \cdot \text{likelihood}}{\text{normalizing constant}} = \frac{p(y^*)L(y^* \mid x_1)}{p(x_1)}$$

- Here, the normalizing constant $p(x_1)$ is:

$$\begin{aligned} p(x_1) &= \sum_{\text{all } y} p(y)L(y \mid x_1) \\ &= p(y = A)L(y = A|x_1) + p(y = C)L(y = C|x_1) + \\ &\quad p(y = G)L(y = G|x_1). \end{aligned}$$

Examples of Calculations

- See R example for table giving counts broken down by species types and weight category.
- Given a penguin that is below average weight, the probability that it is "Adelie" is

$$p(y = A \mid x_1 = 0) = \frac{126}{193} \approx 0.6528$$

- Confirm that this follows Bayes' Rule:

$$p(y = A) = \frac{151}{342}, \quad p(y = C) = \frac{68}{342}, \quad p(y = G) = \frac{123}{342}.$$

$$L(y = A \mid x_1 = 0) = \frac{126}{151} \approx 0.8344$$

$$L(y = C \mid x_1 = 0) = \frac{61}{68} \approx 0.8971$$

$$L(y = G \mid x_1 = 0) = \frac{6}{123} \approx 0.0488$$

Examples of Calculations (continued)

- The normalizing constant is

$$p(x_1 = 0) = \frac{151}{342} \cdot \frac{126}{151} + \frac{68}{342} \cdot \frac{61}{68} + \frac{123}{342} \cdot \frac{6}{123} = \frac{193}{342}.$$

- The posterior probability the penguin is "Adelie" is:

$$\begin{aligned} p(y = A \mid x_1 = 0) &= \frac{p(y = A)L(y = A \mid x_1 = 0)}{p(x_1 = 0)} \\ &= \frac{(151/342) \cdot (126/151)}{193/342} \\ &\approx 0.6528 \end{aligned}$$

- By similar calculations,

$$p(y = C \mid x_1 = 0) \approx 0.3161 \quad \text{and} \quad p(y = G \mid x_1 = 0) \approx 0.0311.$$

Conclusions

- By far, the category with the highest posterior probability is "Adelie".
- Even though the proportion of Chinstraps that are below average weight is more than the proportion of Adelies, the fact the Adelies are much more common in the population than Chinstraps makes it more likely that a random observed below-average-weight penguin is an Adelie.
- Again, this reflects the fact that we set our prior probabilities $p(y = A)$, $p(y = C)$, and $p(y = G)$ to match the species proportions in the sample.
- We could redo the calculations with other prior probabilities (like letting $p(y = A) = p(y = C) = p(y = G) = 1/3$) and the posterior probabilities would be somewhat different.
- But the way we did it is probably the best approach.

Example of Naive Bayes Classification with One Continuous Predictor

- Now let's suppose we want to classify the penguins on the basis of one **continuous** predictor.
- For example, let's classify the penguins on the basis of $X_2 =$ bill length.
- Suppose we observed a penguin with a bill length of 50 mm.
- A plot (see R example) shows that this bill length would be extremely **uncommon** for an Adelie.

Naive Bayes Classification with One Continuous Predictor

- When the predictor is continuous, the naive Bayes approach assumes that the predictor follows a **separate** (conditional) normal distribution for **each level** of the categorical response:

$$X_2 \mid (Y = A) \sim N(\mu_A, \sigma_A^2)$$
$$X_2 \mid (Y = C) \sim N(\mu_C, \sigma_C^2)$$
$$X_2 \mid (Y = G) \sim N(\mu_G, \sigma_G^2)$$

- This is somewhat restrictive, but it's sensible for bill length here, based on the estimated density plots.
- We generally set the means and variances of these normal distributions to equal the sample means and sample variances for the sample data from each species category.

Using Bayes' Rule to Get Posterior Probabilities for each Category

- Then we can find the posterior probability that an observation belongs the category y^* using Bayes' Rule:

$$p(y^* \mid x_2) = \frac{p(y^*)L(y^* \mid x_2)}{p(x_2)} = \frac{p(y^*)L(y^* \mid x_2)}{\sum_{\text{all } y} p(y)L(y \mid x_2)}.$$

- The calculations are (see R code for calculating the heights of the normal densities):

$$p(x_2 = 50) = \frac{151}{342} \cdot 0.0000212 + \frac{68}{342} \cdot 0.112 + \frac{123}{342} \cdot 0.09317 = 0.05579.$$

$$p(y = A \mid x_2 = 50) = \frac{(151/342) \cdot 0.0000212}{0.05579} \approx 0.0002.$$

and similarly,

$pf(y = C \mid x_2 = 50) \approx 0.3992$ and $p(y = G \mid x_2 = 50) \approx 0.600$

Conclusions

- For a penguin with a bill length of 50 mm, the category with the highest posterior probability is "Gentoo".
- Again, the fact that Gentoos are much more common in the population than Chinstraps gives Gentoos an advantage.

Naive Bayes Classification with Two Continuous Predictors

- We can certainly incorporate multiple predictors into the Naive Bayes Classification framework.
- In the penguin example, we can see that including both $X_2 =$ bill length and $X_3 =$ flipper length might improve the classification accuracy (see symbolic scatterplot).
- We can use Bayes' Rule as usual, but in the likelihood part $L(y|x_2, x_3)$, we make the naive (and quite possibly wrong) assumption that X_2 and X_3 are independent, so that

$$L(y \mid x_2, x_3) = f(x_2, x_3 \mid y) = f(x_2 \mid y) f(x_3 \mid y).$$

- In fact, in our penguin example, bill length and flipper length are probably NOT independent; from the scatterplot, note the positive association.

Calculations

Consider a new penguin with bill length $X_2 = 50$ and flipper length $X_3 = 195$:

$$p(y = A)L(y = A \mid x_2 = 50, x_3 = 195) = \frac{151}{342} \cdot 0.0000212 \cdot 0.04554$$

$$p(y = C)L(y = C \mid x_2 = 50, x_3 = 195) = \frac{68}{342} \cdot 0.112 \cdot 0.05541$$

$$p(y = G)L(y = G \mid x_2 = 50, x_3 = 195) = \frac{123}{342} \cdot 0.09317 \cdot 0.0001934$$

$$\sum_{\text{all } y} p(y)L(y \mid x_2 = 50, x_3 = 195) \approx 0.001241.$$

$$p(y = A|x_2 = 50, x_3 = 195) = \frac{\frac{151}{342} \cdot 0.0000212 \cdot 0.04554}{0.001241} \approx 0.0003.$$

Calculations

And similarly,

$$p(y = C|x_2 = 50, x_3 = 195) \approx 0.9944$$
$$p(y = G|x_2 = 50, x_3 = 195) \approx 0.0052.$$

Conclusions

- This penguin is almost certainly a Chinstrap.
- The combination of bill length and flipper length points to a type of penguin that matches the Chinstrap characteristics for this combination of the variables.

Doing It the Easy Way: The `naiveBayes` Function

- To avoid all these tedious calculations, we can simply use the `naiveBayes` function in the `e1071` package in R.
- This calculates the prior category probabilities based on the observed category proportions in the sample (which we said was the preferred approach).
- We can predict the class of a "new" penguin with specified predictor values (see R example)

Assessing the Performance of the Naive Bayes Classification

- Our tools for assessing classification accuracy are similar to those in Chapter 13: The confusion matrix and cross-validation estimates of classification accuracy.
- If we have multiple potential predictor variables, we could build several classification models and compare their performance based on these criteria.
- See R examples for these approaches on the penguins data set.

Naive Bayes versus Logistic Regression

- We have seen that if the categorical response has more than two categories, logistic regression is not an option.
- However, other generalized linear models for multicategory responses exist, though we will not cover them in this class.
- When the response is binary (two categories), there are some advantages to using logistic regression rather than naive Bayes.
- We do get some information from the logistic regression coefficients about the nature of the relationship between the response and the predictors, which we don't get from naive Bayes.
- And naive Bayes makes some simplifying assumptions (normally distributed predictors, independence of predictors) that may not be accurate in reality.
- It's good to know about both of these tools for classification of observations.