Chipy

Naïve Bayes Classifier

* Bayes theorem
* Classification
* How Bayes can be used to classify

One Saturday, I found myself outside a large red brick building in Chicago. Large buildings loomed on either side of the road, with little foot traffic at 3PM. Not a vibrant place at this hour, but instead, a place that made me feel small, as empty corporate buildings towered overhead. I rang the doorbell to a particular building, and was greeted by my Chipy mentor at the door.

My mentor invited me in, and we took the elevator up several flights, eventually emerging into a vestibule on the third story. Opening the doors, I was met with a comfortable space which seemed both old and young at the same time. The floorboards were uneven, showing the property’s age. But inside the building were signs of creativity and energy – phrases painted on the walls, space to relax, eat and work, and conference rooms cleverly named after old bands.

After chatting some to catch up, we got down to business coding. Since my interest in Python is on the Data Science side, we began to discuss a dataset containing album reviews from various critics over the past (x) years (Kaggle link) . The goal: write a program to predict what score a critic would give an album, based on the text of the review.

Which brings me to the topic of this post: Naïve Bayes Classification.

Naïve Bayes Classification

Naïve Bayes Classification really incorporates two concepts: classification, and Bayesian statistics. We will deal with each separately.

Classification: Classification is the problem of trying to predict what category a dataset belongs to. For example, if I could use a classifier to do the following predictions:

|  |  |
| --- | --- |
| **What I want to predict** | **Data to make prediction from** |
| Genre of a book | Text on back cover |
| If a patient is sick | Health data (temperature, etc.) |
| Species of a plant | Flower petal size and stamen size |

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

EASY

For example, imagine we have a camera that takes a picture of a fruit and can tell what color the fruit is. Using the color, we want to guess what type of fruit the camera took a picture of. We took pictures of 100 fruit and the data are below:

|  |  |  |
| --- | --- | --- |
|  | Yellow | Green |
| Apple | 26 | 29 |
| Banana | 30 | 15 |

Now, we let a computer run the camera automatically. In this case, imagine that the camera tells us that the next fruit up is green. How can the computer tell what type of fruit is it?

To answer this question, we will make our first basic fruit classifier. Our classifier will guess that the fruit belongs to the category that shows up most. So, let’s see how many green fruits are apples, and how many are bananas.

The number of green fruits that are apples? 29

The number of green fruits that are bananas? 15

So apples win! Our simple classifier classifies all green fruits as apples.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

MEDIUM

Ok, time to kick it up a notch. Now, in addition to our camera, we have someone measure the fruit size and tell us if the fruit is: Small, Medium or Large. We get a batch of another 100 pieces of fruit and the data are below:

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Color | Size | Fruit |
| 1 | Yellow | Large | Banana |
| 2 | Green | Medium | Apple |
| 3 | Green | Large | Banana |
| 4 | Green | Small | Banana |
| …. |  |  |  |
| 100 | Yellow | Medium | Apple |

So, imagine now that the 101st piece of fruit comes to the system, and we don’t have a human there to check what type of fruit it is. This fruit is green and small. Can we guess what type of fruit it is?

First, I need to figure out the number of Apples that are both green and small (pandas groupby)

Then, I need to calculate the number of Bananas that are green and small.

Finally, I can calculate the probability that my green, small fruit is an apple or a banana.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

HARDER

Classification seems easy! So far, it’s just been about counting, which isn’t so bad. But what happens if we have continuous numbers, instead of colors and size categories? What if we have the fruit diameter in cm?

|  |  |  |
| --- | --- | --- |
| ID | Diameter | Fruit |
| 1 | 11.1 | Banana |
| 2 | 20.8 | Apple |
| 3 | 12.5 | Banana |
| 4 | 10.0 | Banana |
| …. |  |  |
| 100 | 15.2 | Apple |

At this point, the counting method stops being useful. Before, we could describe our dataset simply by counting. Now, we need something else to describe the data. Specifically, we need a probability model.

One of the simplest probability models is the Normal (Gaussian) distribution.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

MEDIUM

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | x | |
|  |  | Yellow | Green |
| Y | Apple | 26 | 29 |
| Banana | 30 | 15 |

x = fruit color

y = fruit type

P(x = yellow) = probability that the fruit is yellow = 56/100

P(x = green) = probability that the fruit is green = 44/100

P(y = apple) = probability that the fruit is an apple = 55/100

P(y = banana) = probability that the fruit is a banana = 45/100.

For example, imagine we have a camera that takes a picture of a fruit and can tell what color the fruit is. Using the fruit color, we want to guess what type of fruit the camera took a picture of. We took pictures of 10 fruit and the data are below:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | x | |
|  |  | Yellow | Green |
| Y | Apple | 11 | 83 |
| Banana | 5 | 1 |

x = fruit color

y = fruit type

P(x = yellow) = probability that the fruit is red

P(x = green) = probability that the fruit is green

P(y = apple) = probability that the fruit is an apple

P(y = banana) = probability that the fruit is a banana.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | x | |
|  |  | Yellow | Green |
| Y | Apple | 5 | 29 |
| Banana | 30 | 15 |

Naïve Bayes is an algorithm that can be used to solve classification problems

Naïve Bayes is based on a mathematical theorem developed by (Reverend Thomas Bayes) in the (x). Bayes’ Theorem relates the probabilities of observing data and category together.

The probability of observing x and y together = P(y|x) \* P(x) = P(x|y)\* P(y)

EASY

For example, imagine we have a camera that takes a picture of a fruit and can tell what color the fruit is. Using the color, we want to guess what type of fruit the camera took a picture of. We took pictures of 100 fruit and the data are below:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | x | |
|  |  | Yellow | Green |
| Y | Apple | 26 | 29 |
| Banana | 30 | 15 |

Now, when we let the camera run automatically, we can use the data above to guess the type of fruit from the color the camera detects.

For example, imagine that the camera tell us that the next fruit up is green.

What is the probability that the green fruit is an apple?

What is the probability that the green fruit is a banana?

x = fruit color

y = fruit type

P(x = yellow) = probability that the fruit is yellow = 56/100

P(x = green) = probability that the fruit is green = 44/100

P(y = apple) = probability that the fruit is an apple = 55/100

P(y = banana) = probability that the fruit is a banana = 45/100.

P(y=apple|x=green) = P(x=green|y=apple)\*P(y=apple)/ P(x=green)

P(y=apple|x=green) = (29/55)\*(55/100)/ (44/100)

P(y=banana|x=green) = P(x=green|y=banana)\*P(y=banana)/ P(x=green)

P(y=banana|x=green) = (15/45)\*(45/100)/ (44/100)