Intro to Data Science:

Naïve Bayes Classification

Ed Podojil Data Engineer/Scientist, Animoto

Warm Up

- Pull latest class git repo, open 'warmup_04.R'
- Comment each line of code, explaining what it does.
- Add a commented summary to the top, explaining what the whole R file does.
- Copy to your personal git repo, commit and push this file to your repo.

Agenda

I. Let's Talk about Probability

II. Naïve Bayesian Classification

III. Implementing a Spam Filter

I. Let's Talk about Probability

Goals

- What is a **probability**
- Different ways to project probability
- How Algebra helps us find Bayesian Inference

What is a probability?

A number between 0 and 1 that determines the likelihood that some event will happen

Representation: We'd say the probability (P) of event A is P(A).

What do we call the space for all possible events?

Sample Space Ω

All events (such as A) are included in this space

Intersecting Probabilities

P(A) == probability of being female

P(B) == probability of having long hair

How do we show the probability of being female & having long hair?

P(A,B)

Conditional Probability

What if we know the "long hair" event happened, and want to know our new probability of A (being female)?

$$P(A|B) = P(A,B) / P(B)$$

or... in layman's terms...

probabiltiy with the help of small animals

Sample Space

$$P(\text{puppy}) = 0.44$$

$P(\text{dark}) = 0.56$

$$P(\text{puppy}|\text{dark}) = P(\text{puppy}, \text{dark}) / P(\text{dark})$$

$$P(\text{puppy}|\text{dark}) = .22 / .56 = \sim.4$$

What about when two events are **independent**?

These two events do not effect each other!

Math way: $P(A|B) = P(A)$

Or another way: $P(A|B) = P(AB) / P(B) = P(A)$

Or even another way: $P(AB) = P(A) * P(B)$

Hang on! Math's almost over!

$P(AB) = P(A|B) * P(B)$ and $P(BA) = P(B|A) * P(A)$

and $P(AB) = P(BA)$

$P(A|B) * P(B) = P(B|A) * P(A)$ and $P(A|B) = P(B|A) * P(A) / P(B)$

Bayes' theorem!

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Cool. What can we do with it?

It's a simple algebraic relationship

It's interesting because it's kind of like a wormhole between two different interpretations of probability

It's very powerful as a computational tool

Wait, two interpretations of probability?

The frequentist interpretation regards an event's probability as its limiting frequency across a very large number of trials.

The Bayesian interpretation regards an event's probability as a "degree of belief," which can apply even to events that have not yet occurred.

Practice

Balls.

An Orange box has 3 balls: 2 white, and 1 black.

A Green box has 3 balls: 2 black, and 1 white.

1. If I were to take a ball from a box, what's the probability that it came from the orange box?
2. If I were to take a white ball from a box, what's the probability that it came from the orange box?

The Theorem again: $P(A|B) = P(B|A) * P(A) / P(B)$

Balls 2.

An Orange box has 6 balls: 5 white, and 1 black.

A Green box has 3 balls: 2 black, and 1 white.

1. If I were to take a ball from a box, what's the probability that it came from the orange box?
2. If I were to take a white ball from a box, what's the probability that it came from the orange box?

The Theorem again: $P(A|B) = P(B|A) * P(A) / P(B)$

Take some time to write some questions down now.
We might answer them in the next section.

(If we don't, then ask during break!)

II. Naïve Bayesian Classification

Goals:

- Understand how our previous formula can be used as a Machine Learning algorithm
- Recognize the different parts of Naïve Bayesian Classification algorithm
- Be able to explain what makes this algorithm naïve

Bayesian Inference

Suppose we have a dataset with n features, and a class label c.

What can we say about classification using Bayes' theorem?

Bayesian Inference

We can use it to help us determine probability of a record belonging to a certain class, given the observations we have.

Review: What is Bayes' Theorem?

Let's look at Bayes' Theorem one more time...

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

And here's our classification algorithm!

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C)\, P(\text{class } C)}{P(\{x_i\})}$$

Notice anything?

Likelihood Function

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C)\, P(\text{class } C)}{P(\{x_i\})}$$

In red we have the **likelihood function**.

It represents the joint probability of all x, given the record belongs to class C

Review: What's a class label?

Review: How do we define joint probability?

The Prior

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C)\, P(\text{class } C)}{P(\{x_i\})}$$

In blue we have the **prior probabilty of c**.

It represents the probability a record belongs to c, regardless of data, as well as while reading in data.

Normalization Constant

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C)\, P(\text{class } C)}{P(\{x_i\})}$$

In green we have the **normalization constant**.

It doesn't depend on class label. We generally ignore it until the end of the computation.

The Posterior

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C)\, P(\text{class } C)}{P(\{x_i\})}$$

In purple we have the **posterior probability of c**.

It represents probabilty AFTER data has been taken account.

Our goal is for the machine to find (or "learn") this distribution, so we can use it for prediction.

{x(i)}: What are we trying to represent here?

{x1, x2, x4, x5, x6...xn}

What's so complicated about this?

Review: What might this kind of look like in programming?

Class break

III. Spam filters

Goals

Preprocess Data

Perform Naïve Bayes Classification

R Reading

- Install and load the E1071 and TM libraries into R
- Find the tm package pdf on CRAN and scan through the documentation. (google search "tm cran documentation")
- Find at least three things/functions that are interesting to you about this package. Write them down and explain what they do.
- Evaluate what we'll likely be using this library for!

Download this file! (type in terminal)

curl \
https://raw.github.com/johnmyleswhite/ML_for_Hackers/master/03-Classification/email_classify.R \
-O email_classify.R

While We Hack:

What are we trying to accomplish?

Prove that Bayes Inference works!

How are we doing that?

Using Bayes Inference on data sets where it will definitely work!

Final Discussion