Bayes Rule

Bayes Rule

- Bayes rule stands out from our other algorithms
- “Bayesian” models don’t so much optimize a mathematical equation as they apply Bayes rule to evidence
- Bayes seems like an outlier among algorithms, but it is deeply connected to linear models

Bayes Rule (cont.)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Lot to unpack here
- P(A|B) means, “The probability of Event A occurs, given Event B occurs”
- P(A) means, “The probability Event A occurs”
- It all seems a little confusing…

Bayes Terms

- P(A|B) is known as the posterior or your “updated” beliefs.
- P(A) is known as the prior and can represent your “initial” (or prior!) beliefs.
- P(B|A) is the likelihood or “evidence.”
- P(B) is a normalizing factor.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} or\ Posterier = \frac{Likelihood}{P(B)} Prior$$

An Example

- A drug test has a true positive rate of 99% and a true negative rate of 97%.
- Surveys indicate that the drug being tested is used by 5% of the population.
- Given a positive test, what is the chance the individual is actually a drug user?
  - Event B = positive drug test
  - Event A = a drug user
  - $P(A|B)$ = What is the probability a positive drug test indicates a drug user?
  - $P(A)$ is 0.05
  - $P(B)$ is $(0.05)(0.99) + (0.95)(0.03)$ (Remember, B is **all** positive tests!)

Example Part II

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B) = \frac{(0.99)(0.05)}{(0.05)(0.99)+(0.95)(0.03)}$$

$$P(A|B) = 0.635$$

DataScience@SMU

Bayes Rule for Multivariables

Multiple Variables and Bayes

$$P(A|B,C) = ?$$

Probability of A given B and C
or let's put B and C in a vector "x"

$$P(A|B,C) = P(A|x) = \frac{P(x|A)P(A)}{P(x)}$$

Probability Math

$$P(A|x) = \frac{P(x|A)P(A)}{P(x)}$$

The numerator is the joint probability.

$$P(x|A)P(A) = P(A, x_1, x_2, ..., x_n)$$

$$P(A, x_1, ..., x_n) = p(x_1|x_2, ..., x_n, A)p(x_2, ..., x_n, A)$$

$$P(A, x_1, ..., x_n) = p(x_1|x_2, ..., x_n, A)P(x_2|x_3, ..., x_n, A)P(x3, ..., x_n, A)$$

$$P(A, x_1, ..., x_n) = P(x_1|x_2, ..., x_n, A)P(x_2|x_3, ..., x_n, A)...P(x_n|A)P(A)$$

Naïve Bayes

- Assume all “x” are independent:

$$P(x_i|x_{i+1},...x_n,A) = P(x_i|A)$$

- Thus:

$$P(A,x_1,...,x_n) = P(x_1|A)P(x_2|A)...P(A)$$

$$P(A|B,C,D) \propto P(A)\prod_{i=1}^{n} P(x_i|A)$$

$$P(A|B,C,D) \propto P(A)P(B|A)P(C|A)P(D|A)$$

What Happened to the Denominator?

- Since we look at classes, the denominator is always the same.
- The denominator is also difficult to calculate.
  - Numerator is based on evidence—easily measured
- So if we are looking at classes, we take the class with the max score!
  - Before “A” was 1 class, but what if it were 2 classes (R, S)?

$$P(R|B,C,D) \propto P(R)P(B|R)P(C|R)P(D|R)$$

$$P(S|B,C,D) \propto P(S)P(B|S)P(C|S)P(D|S)$$

We pick class R or S based on the higher score! (argmax)

DataScience@SMU

Bayes Rule for Continuous Variables

Continuous Variables

- So far we have talked of events and classes.
- These are all categorical variables.
- How do we deal with “continuous” data like temperature or weight?

Statistics

- Using statistics, we can represent the probability that a value came from certain distributions.
- Example: If we measured the weight of all the people in the United States, we could model that with a normal distribution with a mean and standard deviation.
  - Then, if we measured an unknown person's weight, we could calculate the probability that person was from the United States.
  - Gaussian density:
    - x is your measurement
    - μ is the mean
    - σ is the standard deviation

$$\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

DataScience@SMU

Naïve Bayes and Logistic Regression

Naïve Bayes and Logistic Regression

- Recall that to classify R or S, we looked at the max value of the following:

$$P(R|B,C,D) \propto P(R)P(B|R)P(C|R)P(D|R)$$

$$P(S|B,C,D) \propto P(S)P(B|S)P(C|S)P(D|S)$$

- We can rewrite this for class R as:

$$\frac{P(R|B,C,D)}{P(S|B,C,D)} > 1$$

- We can rewrite this for class S as:

$$\frac{P(R|B,C,D)}{P(S|B,C,D)} < 1$$

The Logit

- Logit is defined as:

$$logit(p) = log_e(\frac{p}{1-p}) = log_e(p) - log_e(1-p)$$

- Assume Class R and S is a binary classifier. Then p(S) = 1 – p(R)

$$log_e(\frac{P(R|x)}{P(1-R|x)}) > 0$$

X here is just the vector of features and log 1 = 0

Log Loss and Bayes Rule

- Logistic regress optimizes a linear equation for the log loss equation.
- Naïve Bayes estimates the loss based on the data.
- The log loss turns out to be Bayes rule!
- The two classifiers form a generative-discriminative pair:
  1. Generative: Bayes
     - Reaches asymptotic error faster
  2. Discriminative: logistic regression
     - Can have lower asymptotic error

DataScience@SMU