NAÏVE BAYES CLASSIFICATION

Paul Speaker

Review of Bayesian Analysis
Covid Testing Example

- Suppose 3% of a population has Covid. A Covid test is 95% accurate when the subject has Covid and is 90% accurate when the subject does not have Covid. If a test comes up positive, what is probability the person has Covid?
  - **Not 95%!!!!**
  - Note: this quantity is called the Positive Predictive Value (PPV)
- 2 ways a test can be positive
  - Person has Covid, and the test is right (true positive)
  - Person does not have Covid, and test is wrong (false positive).
- P(Covid) = N(A)/N = P(true positive) / (P(true positive) + P(false positive))
  - P(true positive) = 0.95 x 0.03
  - P(false positive) = 0.97 x 0.10
- A = has Covid
- B = positive test
- **Example of Bayes Theorem**

Likelihood

Prior

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

Bayes Theorem

- Recall that output of any classification model is a conditional probability
  - P(Y = y | X = x): Probability that Y is in a class, given that the x's are what they are
- We can use the reverse case through Bayes Theorem
  - P(X = x | Y = y): Probability that the x's are what they are, given that Y is in a class
  - These statistics are straightforward to calculate, based on the data
  - For discrete X's, we can use proportions
  - For continuous X's, we can use pdf's
- For naïve Bayes, two assumptions will be made
  - Input variables are independent of each other
  - Continuous input variables follow a normal distribution
  - Both of these assumptions can be modified for more sophisticated models

Naïve Bayes with Discrete Input

- Below is a very simple example using Bayes Theorem with 1 categorical input
- Suppose we are predicting the probability that Y = 1, given that X = M, with data summarized in the table below

	X = S	X = M	X = L
Y = 0	13	21	33
Y = 1	19	37	22

- The use of Bayes Theorem to predict the probability is as follows, with
- P(Y = 1) = 78/145 (prior)
- P(X = M | Y = 1) = 37/78
- P(Y = 0) = 67/145
- P(X = M | Y = 0) = 21/67

$$P(Y=1|X=M) = \frac{P(Y=1)P(X=M)|Y=1)}{P(X=M)}$$

$$= \frac{P(Y=1)P(X=M)|Y=1)}{P(Y=1)P(X=M|Y=1)+P(Y=0)P(X=M|Y=0)}$$

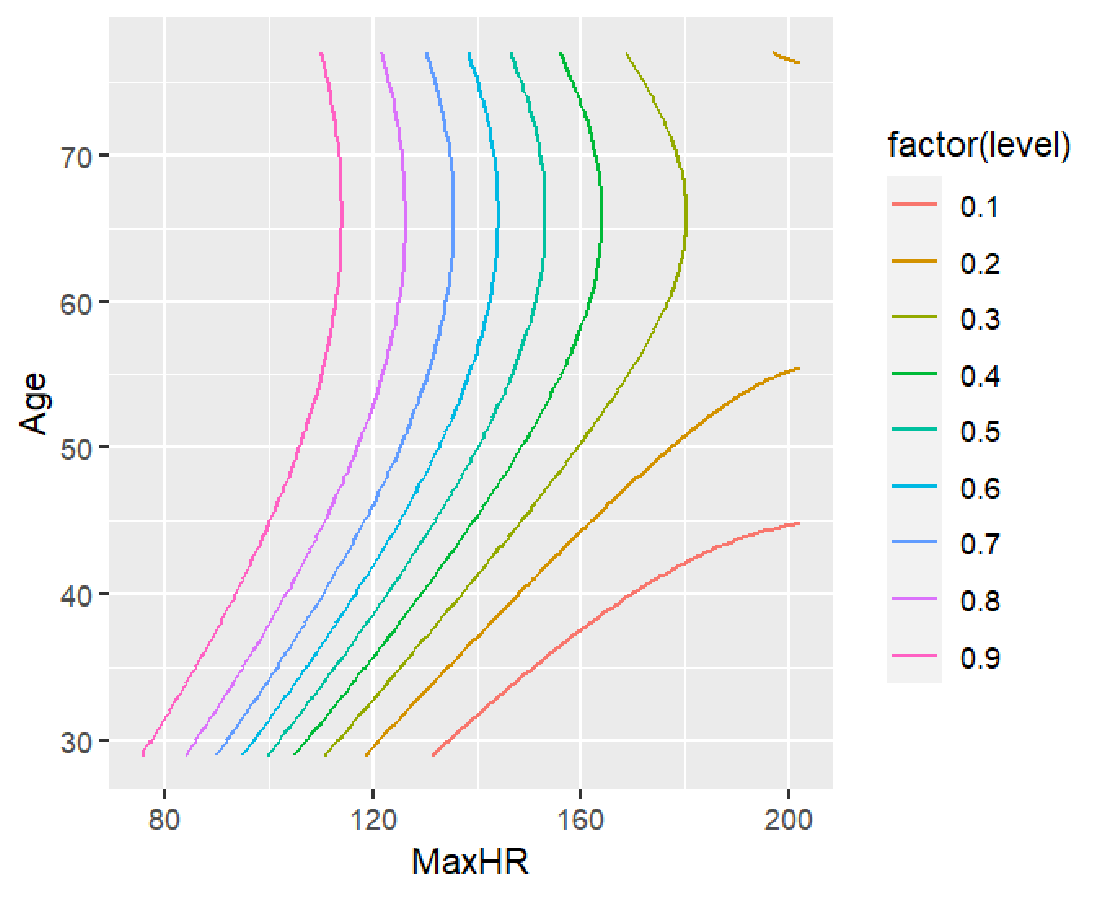
- The math yields P(Y = 1 | X = M) = 37/58 (which is easy to see in the data)

Naïve Bayes with Continuous Input

- When a continuous input is used for Bayes Theorem the pdf replaces the proportions (the proportions are estimates of pmf)
- Naïve Bayes assumes a normal distribution for X's, and that the X's are independent of each other
  - Separate normal distributions for X based on target value
  - Parameters are estimated for each class based on mean and standard deviations
  - Parameter count: if there are 2 target classes, each X has 4 parameters!
    - More complex than logistic regression, which only has 1 parameter per X
- To do naïve Bayes estimate, we can use dnorm for continuous input
- When there are multiple continuous inputs, pdf is product of normal distributions (independence)
- When there are both continuous and discrete inputs, pdf is product of normal distribution(s) and proportions for discrete (still independence)

Naïve Bayes Implementation in R

- Naïve Bayes is straightforward enough to calculate directly. Steps:
  - For each class, calculate
    - The prior (proportion in entire dataset)
    - The likelihood
      - Proportions for categorical
      - Mean and standard devisions for dnorm
      - Likelihood is products of dnorm and proportions
- Plug into Bayes' Theorem
- Also can do with naivebayes command
  - Requires e1071 library



Advantages and Disadvantages of Naïve Bayes

- Advantages:
  - Simple to implement
  - Computationally not costly (more parameters, but easier to fit)
    - Parameter fit does not involve optimization!
    - Often fastest option for large datasets
  - Can handle imbalanced datasets better than most
    - Higher uncertainty for parameters for smaller class
- Disadvantages
  - Independence Assumption
    - Can fix for numerical inputs with PCA (independent linear combinations of X's)
  - Normality assumption
  - Number of parameters means more likely to overfit
    - Particularly a problem with multivalued classification

Beyond Naïve Bayes

- Some of the disadvantages can be handled with model modifications
  - Normality for continuous variables
    - Modifications of naïve Bayes exist to use non-normal distributions
    - Can use kernel density estimate
      - Non-parametric approach
      - pdf of kde is smoothed histogram for data
  - Independence
    - Rather than independent normal variables, can take covariance into account
      - Linear Discriminant model (next time)
        - Solve independence assumption for continuous only
        - Other problems
      - Handling dependence with categorical variables is harder but doable
        - For categorical/categorical dependence, use proportions for each combination
        - For categorical/continuous dependence, calculate statistics of continuous variables for each input/target class combination