NAÏVE BAYES CLASSIFICATION

Paul Speaker

Bayesian Models

- Bayesian Classification Models make use of Bayes Theorem to calculate conditional probabilities.
- The only difference is with the likelihood calculation
- Naïve Bayes → X independence
- Other Bayesian models allow for covariance
- Simplest type—Linear Discriminant Model

Likelihood

Prior

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

The Covariance Matrix

- The Covariance Matrix (sometimes called the Variance/Covariance Matrix) is a fundamental tool for advanced data science methods
- For a joint distribution of 2 variables, the covariance matrix is given as

$$\Sigma = \begin{bmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{bmatrix}$$

- Clearly Symmetric
- The fact that correlation is between -1 and 1 yields the important fact that the covariance matrix is positive definite
  - Complete set of (real, positive) eigenvalues and (orthogonal) eigenvectors
  - Eigenvectors of the Variance matrix are called the ***Principal Components***

mvnorm Functions in R

- A ***multivariate normal random variable*** is a joint distribution where the marginal distributions are normal variables
- The pdf of a multivariate random variable is specified by a mean value (mu) and a covariance matrix (Sigma)
- Functions in the MASS and mvtnorm libraries to work with multivariate normal random variables
- From mvtnorm, we can use dmvnorm(x, mu, Sigma) to find the density at a point x
  - Dmvnorm can be used as likelihood function to accommodate covariances (in LDA or QDA)
- From MASS, we can a;sp use mvrnorm(n, mu, Sigma) to generate n simulations of the random variable
  - Output is n x 2 for a 2-variable normal random variable

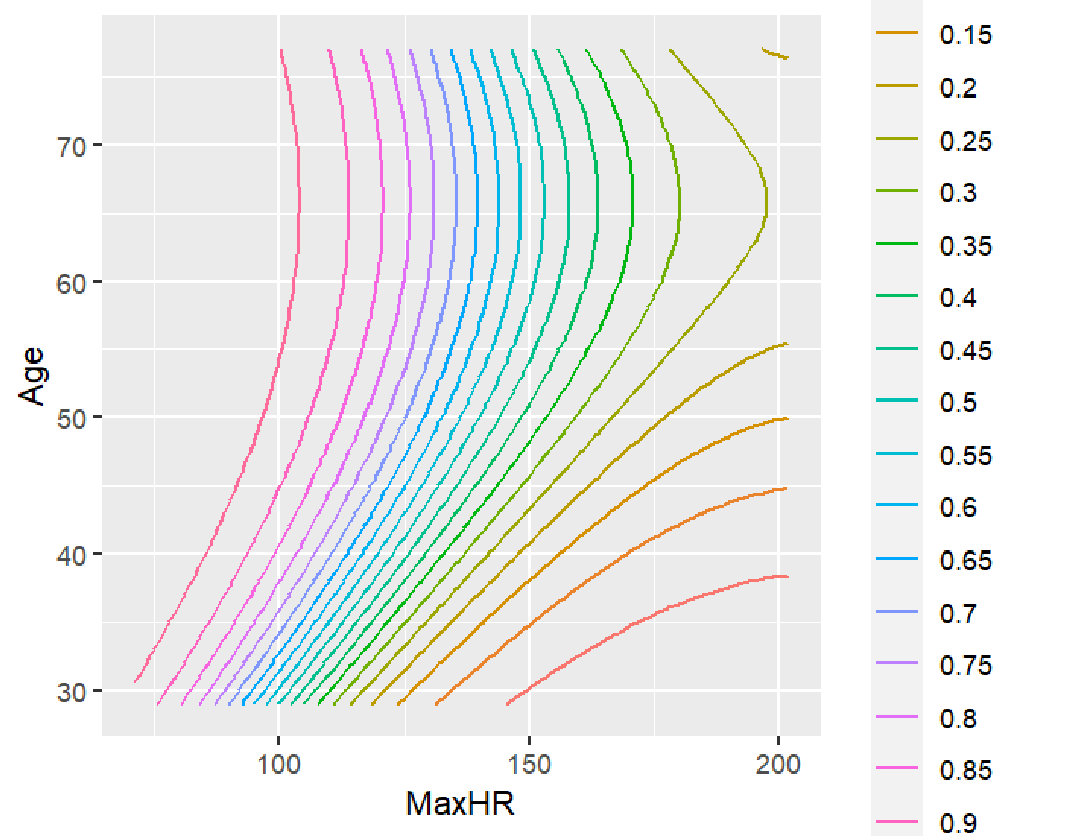
Linear Discriminant Models

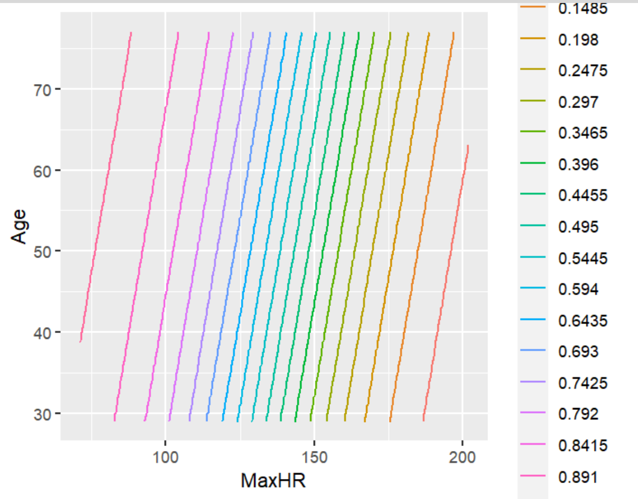
- Types:
  - Linear Discriminant Model (LDA)
  - Quadratic Discriminant Model (QDA)
- LDA is simultaneously more, equally, and less flexible than Naïve Bayes
  - More flexible: allows covariance
  - Equally flexible: continuous input variables assumed to be normally distributed
    - Multivariate Normal variable
  - Less flexible: variances treated as equal across classes
  - Also less flexible: No categorical inputs
- Parameter counts: assuming binary target and n inputs
  - Each class will have n mean values for inputs (2n total)
  - Overall covariance matrix will be size n(n + 1)/2
  - Sizes of covariance matrices are important for larger number of n's

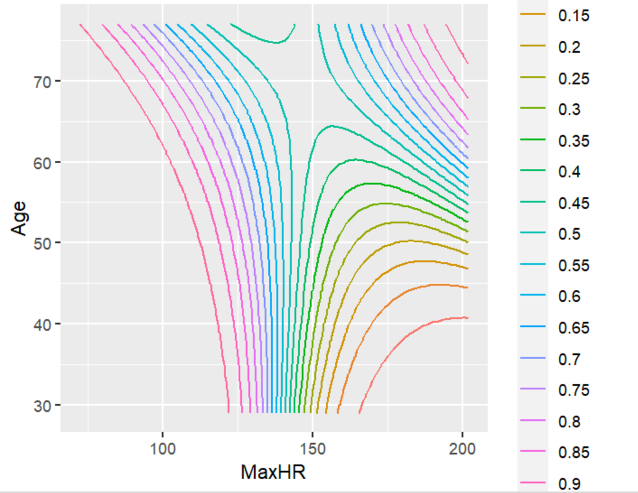
Quadratic Discriminant Models

- A quadratic discriminant model (QDA) is another statistical method used for classification
- Like LDA, it only uses continuous inputs, which it assumes are normally distributed for each class
- Unlike LDA, it uses different covariance matrices for each class.
- Will have ~ twice as many parameters as LDA (same number of means, but twice as many covariance matrix entries)
- For single input, quadratic discriminant model is identical to Naïve Bayes
- Modeling in R is with the MASS library
  - lda(data = <>, y ~ x1 + x2 + …)
  - qda(data = <>, y ~ x1 + x2 + …)

Comparison of Naïve Bayes, LDA, and QDA







Class Covariance Matrix EDA

- Analyzing the Class Covariance Matrices, in relation to the mean values is useful
  - Standard deviations small in comparison to the mean differences → separation between classes, easy to tell
    - Parameters will be unstable for logistic regression
    - But if it's easy to tell, do you care about the parameters?
- Differences between class covariance matrices points to LDA will struggle
- In general, there is little reason to use LDA over QDA, unless large number of inputs (too many parameters from covariance matrices)
- Like Naïve Bayes, LDA and QDA are calculated off of summary statistics, and so are easy to work with
  - Independence can be fixed with PCA, while differences in class covariances cannot be fixed
- Also, how do covariances compare to variances in matrix? (difficult to compare—can be differently scaled, so correlation matrix might be easier)
  - X's can be re-scaled to have similar shape (more on this next time!)

Class Covariance Matrix EDA

- There are many possible combinations for class covariance matrices
- Different combinations→ different models might be better