Naive Bayes

Hunter Glanz

OUTLINE

Recall Our Classification Goal...

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- This measures the training mis-classification or *error* rate

Recall Our Classification Goal...

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- This measures the training mis-classification or *error* rate
- The *test error* rate associated with a set of test observations of the form (x_0, y_0) is

$$\text{Ave}(I(y_0 \neq \hat{y}_0))$$

Minimizing the Test Error

- The classifier that minimizes the test error rate is...

Minimizing the Test Error

- The classifier that minimizes the test error rate is...

The Bayes Classifier

Minimizing the Test Error

- The classifier that minimizes the test error rate is...

  The Bayes Classifier

- ...which *assigns each observation to the most likely class, given its predictor values*!
- In other words, assign class j if

$$\Pr(Y = j|X = x_0)$$

is the largest

Where We're Going

- Logistic regression directly modeled our conditional probabilities
- Some issues with logistic regression:
    1. When classes are well-separated, parameter estimates are surprisingly unstable
    2. Typically n needs to be large
    3. There are simpler methods when it comes to more than 2 classes

Where We're Going

- Logistic regression directly modeled our conditional probabilities
- Some issues with logistic regression:
  1. When classes are well-separated, parameter estimates are surprisingly unstable
  2. Typically n needs to be large
  3. There are simpler methods when it comes to more than 2 classes
- Bayes' Theorem:
  - Model distribution of predictors X separately in each response class
  - Use Bayes' Theorem to flip these into estimates for $\Pr(Y = k | X = x)$

The Math To Start Us Off

- Suppose the response variable has K classes
- Let π_k be the *prior* probability that a randomly chosen observation comes from the kth class, $\Pr(Y = k)$
- Let $f_k(x) = \Pr(X = x | Y = k)$
  - So, $f_k(x)$ is relatively large if there's a high chance an observation in the kth class has $X \approx x$

The Math To Start Us Off

- Suppose the response variable has K classes
- Let π_k be the *prior* probability that a randomly chosen observation comes from the kth class, $\Pr(Y = k)$
- Let $f_k(x) = \Pr(X = x | Y = k)$
  - So, $f_k(x)$ is relatively large if there's a high chance an observation in the kth class has $X \approx x$
- Bayes' Theorem:

$$\Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\Pr(X = x)}$$

- Simple enough, right?

Broadening the Problem a Bit

$$\Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\Pr(X = x)}$$

Broadening the Problem a Bit

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\Pr(X = x)}$$

- In practice, we're only interested in the numerator because the denominator does not depend on k
- The numerator is equivalent to $\Pr(Y = k, X)$ (i.e. the joint probability)

Broadening the Problem a Bit

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\Pr(X = x)}$$

- In practice, we're only interested in the numerator because the denominator does not depend on k
- The numerator is equivalent to $\Pr(Y = k, X)$ (i.e. the joint probability)
- If we have multiple predictors then it's $\Pr(Y = k, X_1, \ldots, X_p)$
- Usually, estimating/computing $\Pr(Y = k, X_1, \ldots, X_p)$ is hard

So what do we do?

Naive Bayes

1. The “naive” part refers to **the assumption that the value of a particular feature (predictor) is independent of the value of any other feature, given the class (response) variable**.

Naive Bayes

1. The "naive" part refers to **the assumption that the value of a particular feature (predictor) is independent of the value of any other feature, given the class (response) variable**.
2. This means:

$$\Pr(Y = k | X_1, \ldots, X_p) \propto \Pr(Y = k) \prod_{i=1}^{p} \Pr(X_i = x_i | Y = k)$$

- In the end, $\Pr(Y = k | X_1 = x_1, \ldots, X_p = x_p)$ is the *posterior* probability that an observation belongs to the kth class, and we usually want to maximize it!

Sound Familiar?!

- This should feel reminiscent of LDA and QDA!

Sound Familiar?!

- This should feel reminiscent of LDA and QDA!
  - Both LDA and QDA make the assumption that the probability distribution of the features is Normal (Gaussian) and...

Sound Familiar?!

- This should feel reminiscent of LDA and QDA!
  - Both LDA and QDA make the assumption that the probability distribution of the features is Normal (Gaussian) and...
    1. LDA assumes that the covariance matrix is the same across all classes.
    2. QDA allows the covariance matrix to vary, but does NOT assume that the features are independent. That is, Gaussian Naive Bayes would involve covariance matrices that are **diagonal**.

Sound Familiar?!

- This should feel reminiscent of LDA and QDA!
  - Both LDA and QDA make the assumption that the probability distribution of the features is Normal (Gaussian) and...
    1. LDA assumes that the covariance matrix is the same across all classes.
    2. QDA allows the covariance matrix to vary, but does NOT assume that the features are independent. That is, Gaussian Naive Bayes would involve covariance matrices that are **diagonal**.
- In any case, to a large degree the biggest work in Naive Bayes is the estimation of the probability distribution!
  - There are many ways to do this that we won't talk extensively about, but there are some popular choices...

Some Discussion

- The independence ("naive") assumption can often be inaccurate, but this classification method can still be quite successful!
- Construction and estimation of this model is often
  1. VERY fast
  2. and does NOT require large amounts of data!