Applied Bayesian Method HW1

Ting-Xuan Wang 112352015

October 7, 2024

1 Assignment Description

In accordance with the framework established by Ng and Jordan (2001), I utilize the breast cancer dataset from the UCI Machine Learning Repository to compare the prediction performance and asymptotic error of discriminative and generative learning approaches. Specifically, logistic regression is employed for the discriminative learning method, while Naive Bayes is used for the generative learning approach.

2 Methodology

2.1 Naive Bayes

Naive Bayes classifiers are grounded in **Bayes' Theorem**, which relates the conditional and marginal probabilities of random variables:

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

where $P(A|B)$ is the probability of event A given that B is true, $P(B|A)$ is the probability of event B given that A is true, $P(A)$ is the probability of event A and $P(B)$ is the probability of event B. The core concept of Naive Bayes is that it assumes the features used for classification are independent given the class label. This assumption simplifies the computation of the joint probability of the features. Mathematically, for a given class C_k and features $X_1, X_2, ..., X_n$, the assumption can be expressed as:

$$P(X_1, X_2, ..., X_n|C_k) = P(X_1|C_k) \cdot P(X_2|C_k) \cdot ... \cdot P(X_n|C_k) \tag{1}$$

Moreover, there are several variants of Naive Bayes, tailored to different types of data. Firstly, the **Gaussian Naive Bayes** assumes that the features follow a Gaussian distribution and it is often used for continuous data. Second, **Multinomial Naive Bayes** is suitable for discrete data, particularly for text classification tasks, where the features represent the frequency of the terms. Lastly, which is similar to Multinomial Naive Bayes, is the **Bernoulli Naive Bayes**, it assumes binary features.

2.2 Logistic Regression

Logistic regression is primarily used for binary classification problems, where the dependent variable can take on two possible outcomes. Instead of predicting a continuous outcome, logistic regression estimates the probability that a given input belongs to a particular category. Mathematically, logistic regression model uses the logistic function (sigmoid function) to map any real-valued number into the range of 0 to 1:

$$f(z) = \frac{1}{1 + e^{-z}}$$

where z is the linear combination of the input features:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Using the logistic function, the predicted probability $P(Y = 1|X)$ for the positive class can be expressed as:

$$P(Y = 1|X) = f(z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}}$$

3 Data

The data we use is the "Congressional Voting Records" dataset from UCI machine learning repository. This dataset includes voting records for 435 members of Congress on 16 different voting issues, and the target variable indicates the party affiliation of the Congress member, which can be either " Democrat" or "Republican". Following table shows additional variable information:

Class Name	2 (democrat, republican)
handicapped-infants	2 (y, n)
water-project-cost-sharing	2 (y, n)
adoption-of-the-budget-resolution	2 (y, n)
physician-fee-freeze	2 (y, n)
el-salvador-aid	2 (y, n)
religious-groups-in-schools	2 (y, n)
anti-satellite-test-ban	2 (y, n)
aid-to-nicaraguan-contras	2 (y, n)
mx-missile	2 (y, n)
mmigration	2 (y, n)
synfuels-corporation-cutback	2 (y, n)
education-spending	2 (y, n)
superfund-right-to-sue	2 (y, n)
crime	2 (y, n)
duty-free-exports	2 (y, n)
export-administration-act-south-africa	2 (y, n)

Since our features is mostly in discrete form, we will use the Bernoulli Naive Bayes for prediction.

4 Result

To compare Naive Bayes classifier and Logistic Regression, I calculate the mean error of both algorithm over 1000 random train/test split, with the training size m becomes larger and larger. The setting allow us to compare not only the error but also the asymptotic error. I let the test data be fixed first and randomly select different size of train size. Next, I let the test data and train data be randomly selected every time I do the prediction. The results are shown in fig (1) and fig (2) respectively.

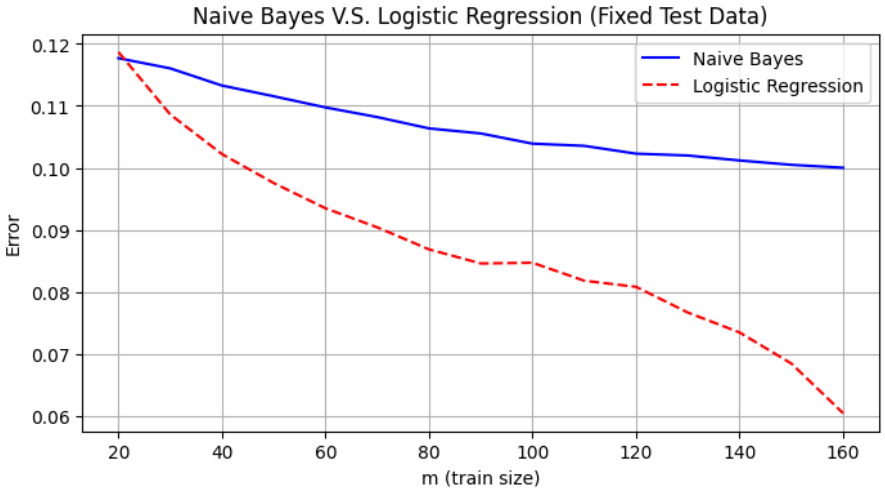

Figure 1: Voting Records (Discrete), Fixed testing data

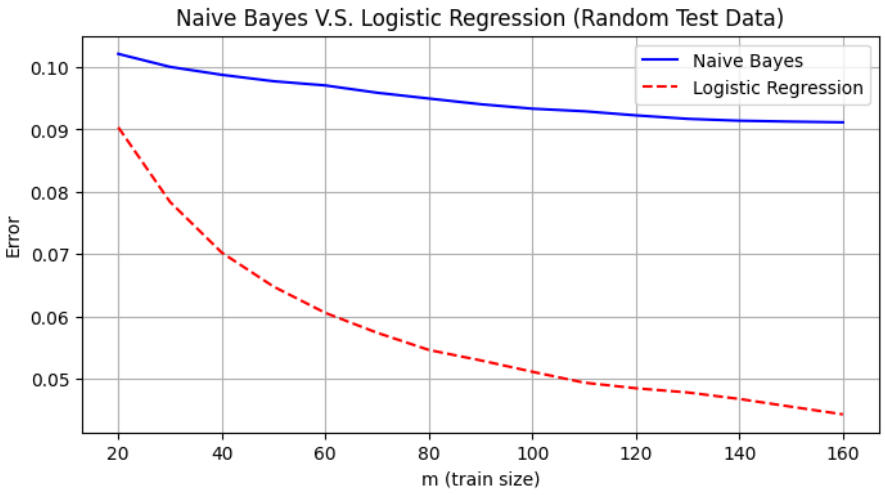

Figure 2: Voting Records (Discrete), Random testing data

The two figures reveal a similar trend: the generative learning approach (Naive Bayes) converges to its asymptotic error much more quickly, whereas the discriminative learning approach (Logistic Regression) achieves a lower asymptotic error. This outcome aligns with the findings of Ng and Jordan (2001).