CS 725: Foundations of Machine Learning (Autumn 2023)— Homework 3 Exploring Classification using Naive Bayes

Anuj Asati 23M0763
Frederic J Maliakkal 23M0745

September 2023

Exploring Classification using Naive Bayes

Naive Bayes is a powerful probabilistic classification algorithm known for its simplicity and efficiency in various machine learning tasks, especially in the fields of natural language processing and text classification. Naive Bayes is based on Bayes' theorem, which is one of the original concepts of probability theory. The algorithm is called "naive" because it assumes a high degree of independence of the features used for classification.

The nature of the data is the determining factor in the choice of the variant to be used. Naive Bayes uses probability theory as a flexible and efficient classification algorithm to generate predictions. It often provides competitive results and is a valuable tool in the machine learning toolbox, although its "naive" assumption of independence may not always hold for real data.

Bayes Theorem

The fundamental theorem in probability theory and statistics is known as Bayes' theorem. It offers a technique to adjust the probability of a hypothesis (an occurrence or statement) in light of fresh data or information. Machine learning, statistics, and Bayesian inference are just a few of the disciplines that make heavy use of the Bayes theorem.
The formula is as follows:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{1.1}$$

Where:

P($A|B$) is the conditional probability of event A occurring given that event B has occurred (the posterior probability).
P($B|A$) is the conditional probability of event B occurring given that event A has occurred (the likelihood).
P(A) is the prior probability of event A (our initial belief in A before considering

evidence).
P(B) is the marginal probability of event B (the overall probability of observing B).

Approach: We are using MLE for parameter estimation and then we calculate the likelihood for every variable in the new test input and calculate the probability of different classes using the learned parameters. We multiply all the probabilities together to get a collective likelihood for all the input data, then we multiply the likelihood with prior info for calculating posterior for every class and use argmax for predicting the class.

Gaussian distribution

The Gaussian distribution is a continuous probability distribution that plays a central role in probability theory and statistics.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1.2}$$

Approach: When we use MLE for distribution, the parameters we end up calculating are mean and variance. We then use the estimated mean and variance to calculate the likelihood using the equation(1.2).

Class	X1		X2	
	Mean	Variance	Mean	Variance
0.0	2.0209	3.9067	9.0528	78.4361
1.0	0.02136	0.8559	25.1634	230.0548
2.0	8.0248	-0.0216	35.6724	4.0079

Bernoulli Distributions

The Bernoulli distribution is a discrete probability distribution that models a random experiment with two possible outcomes: success and failure.

$$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} \tag{1.3}$$

Approach: The MLE estimation calculates the probability of success as the parameter for Bernoulli distribution. we use that estimated parameter as p in the equation(1.3) and calculate the likelihood for finding posterior.

Class	X3_Mean	X4_Mean
0.0	0.2023	0.104
1.0	0.5984	0.8018
2.0	0.9053	0.1947

Laplace Distributions

The Laplace distribution, also known as the double-exponential distribution, is a continuous probability distribution that is characterized by its peakedness around its mean and its heavy tails.

$$f(x;\mu,b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{1.4}$$

Approach: For Laplace distribution we get two parameters when we use MLE, namely Location parameter(μ) and Scale parameter(b). We used median for as it was giving more accuracy than mean. Scale parameter was calculated using

$$b_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$$

After estimating the parameters, those were used to calculate the likelihood and thereby predict the class for new test input data.

Class	X5		X6	
	Median	Mean	Median	Mean
0.0	0.0766	0.8728	1.9835	5.9781
1.0	0.3828	0.3513	0.9993	5.9983
2.0	0.7963	0.2125	3.0050	3.0614

Exponential Distribution

The exponential distribution is a continuous probability distribution that models the time between events in a Poisson process, where events occur continuously and independently at a constant average rate.

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{1.5}$$

Approach: The parameter estimated when we use MLE is lambda(λ) which is the rate parameter calculated as

$$\lambda_{\text{MLE}} = \frac{1}{\text{sample mean}} \tag{1.6}$$

Class	X7_Mean	X8_Mean
0.0	1.9782	3.9354
1.0	2.9841	7.9800
2.0	8.9427	14.6849

Multinomial Distributions

The multinomial distribution is a probability distribution used in statistics to model experiments or trials where there are more than two mutually exclusive outcomes. It is an extension of the binomial distribution, which deals with two possible outcomes (success and failure).

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} \tag{1.7}$$

Approach: For every class in the multinomial distribution we find the frequency of the class and store that in a 1D array as it acts as parameter for multinomial class for MLE.

$$p_{i,\text{MLE}} = \frac{\text{number of observations in category } i}{n}$$

For calculating the likelihood we take the probability of the class in which the input variable belong.for example, if the input is of class 3 then we take the probability of class 3 which is P_3

Class	X9_Probabilities	X10_Probabilities
0.0	0.2022 0.2032 0.2042 0.1967 0.1937	0.1213 0.1236 0.1257 0.1277 0.1270 0.1271 0.1241 0.1235
1.0	0.0977 0.1984 0.4047 0.1583 0.1409	0.1009 0.0506 0.0508 0.1998 0.1524 0.1487 0.2003 0.0965
2.0	0.2052 0.2997 0.1029 0.3417 0.0505	0.1972 0.0481 0.0483 0.1054 0.1552 0.1530 0.0980 0.1948

Priors

Priors play a crucial role in the Bayesian framework, as they are combined with observed data to calculate a posterior probability distribution, which represents your updated beliefs after considering the evidence. In Bayesian probability theory, a prior is a probability distribution that represents your beliefs or uncertainty about a certain parameter or hypothesis before observing any data. It expresses your initial knowledge or assumptions about the parameter or hypothesis.

Class	Priors
0.0	0.3333
1.0	0.3333
2.0	0.3333

Results

Accuracy

Metric	Value
Training Accuracy	0.9014
Validation Accuracy	0.9023

Table 2.1: Model Performance

F1 Score

Class	Training	Validation
0	0.881	0.881
1	0.879	0.878
2	0.943	0.947

Table 2.2: F1 Scores