CS 725: Foundations of Machine Learning Homework 3

Soumen Kumar Mondal
23m2157@iitb.ac.in

Naay Balodia
23m2166@iitb.ac.in

September 17, 2023

Abstract

In this assignment, we will implement Naive Bayes model on a toy datasets for 3 class classification task. Our key goal in this assignment is to correctly implement these models and analyze the results we obtained.

1 Naive Bayes Classifier

In the classification problem, the features ($\bar{x}$) will be given and we are interested to find out what is the probability that a label ($\hat{y}$) will be correctly classified? More generally, given the features, what is the most likely label?

In this homework, it is given that $\bar{x} = [x_1, x_2, \ldots, x_{10}]^T$ that means there are 10 features in the dataset. The class label y can be any one of the class between $[0, 1, 2]$. Probability of the label y given the feature $\bar{x}$ is denoted as $P[y|\bar{x}]$. In this homework problem, since there are 3 classes, we will be interested to know the vector of posterior probabilities of labels for all the classes as:

$$P[y = 0|\bar{x}], P[y = 1|\bar{x}], \ldots, P[y = 2|\bar{x}] \tag{1}$$

The output of the classifier will be given as:

$$\hat{y} = \underset{k=0,1,2}{\text{argmax}} P[y = k|\bar{x}] \tag{2}$$

The posterior probability of label term can be written in terms of prior probability of label and likelihood of features given a label by Bayes theorem as follows:

$$\text{for } k = 0, 1, 2: \quad P[y = k|\bar{x}] = \frac{P[\bar{x}|y = k]}{P[\bar{x}]} \quad \text{(Bayes Theorem)} \tag{3}$$

The denominator term $P[\bar{x}]$ doesn't depend on the label y. Therefore, we can even ignore this term as we are more interested in the likelihood or a score rather than the valid probability measure. In the numerator, the likelihood term can be expanded for all the features that are conditionally dependent on other features and label by multiplicative rule of conditional probability as follows:

$$P[\bar{x}|y = k] = P[x_1 \cap x_2 \cap \cdots \cap x_{10}|y = k] \tag{4}$$

$$= P[x_1|y = k] \cdot P[x_2|x_1 \cap y = k] \ldots P[x_1 0|x_1 \cap x_2 \cap \cdots \cap x_9 \cap y = k] \tag{5}$$

In general, the features are not independent of each other but given a label, it is naively assumed that the features become conditionally independent under the given label. Hence the above equation can be simplified as follows:

$$P[\bar{x}|y=k] = P[x_1|y=k] \cdot P[x_2|y=k] \dots P[x_{10}|y=k] \tag{6}$$

$$= \prod_{i=1}^{10} P[x_i|y=k] \tag{7}$$

$$P[y=k|\bar{x}] = \left(\prod_{i=1}^{10} P[x_i|y=k] \right) \cdot P[y=k] \tag{8}$$

$$\hat{y} = \underset{k=0,1,2}{\operatorname{argmax}} \left(\prod_{i=1}^{10} P[x_i|y=k] \right) \cdot P[y=k] \tag{9}$$

Since the likelihood values are very small, we take log likelihood to avoid numerical underflow. Therefore, the predicted label is given by the following equation[1]:

$$\hat{y} = \underset{k=0,1,2}{\operatorname{argmax}} \left[\sum_{i=1}^{10} (\log(P[x_i|y=k])) + \log(P[y=k]) \right] \tag{10}$$

2 Maximum Likelihood Estimate

In the Equation 10, we have to find the values of $P[x_i|y=k]$ which can be found from the conditional density or mass of the probability distribution corresponding to each of the features. Since the probability distribution of each of the features are known prior, the training set will be used to estimate the parameters of the PMF or PDF using the MLE principle.

For each of the class label y, the dataset is filtered and the parameters are estimated corresponding to filtered dataset. This gives the parameters of PMF or PDF corresponding to a particular class label y. Once the parameters are estimated, based on an unknown feature $\bar{x}$, the class label is predicted from Equation 10. The prior probability is calculated by the frequency of observing a class label in the training dataset. The features and its PMF or PDF is shown in Table 1. The MLE estimated parameters are shown in Table 2.

Features	Distribution	PMF or PDF ($p_X(x)$)
X_1, X_2	$\sim Normal(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(x-\mu)^2}{2\sigma^2})$
X_3, X_4	$\sim Bernoulli(p)$	$p^x(1-p)^{1-x}$
X_5, X_6	$\sim Laplace(\mu, b)$	$\frac{1}{2b} \exp(-\frac{\lvert x-\mu \rvert}{b})$
X_7, X_8	$\sim Exponential(\lambda)$	$\lambda \exp(-\lambda x)$
X_9, X_{10}	$\sim Multinomial([p_1, p_2, \dots, p_k])$	$p_{x-1}(x \epsilon [0, 1, 2, \cdots, k-1])$

Table 1: Features and its PMF or PDF

Features	Distribution	MLE estimated parameters
X_1, X_2	$\sim Normal(\mu, \sigma^2)$	$\hat{\mu} = \frac{\sum_{i=0}^{N-1} x_i}{N}$, $\hat{\sigma^2} = \frac{\sum_{i=0}^{N-1}(x_i - \hat{\mu})^2}{N}$
X_3, X_4	$\sim Bernoulli(p)$	$\hat{p} = \frac{\sum_{i=0}^{N-1} x_i}{N}$
X_5, X_6	$\sim Laplace(\mu, b)$	$\hat{\mu} = median(x_0, x_1, \cdots, x_{N-1})$, $\hat{b} = \frac{\sum_{i=0}^{N-1} \lvert x_i - \hat{\mu} \rvert}{N}$
X_7, X_8	$\sim Exponential(\lambda)$	$\hat{\lambda} = \frac{N}{\sum_{i=0}^{N-1} x_i}$
X_9, X_{10}	$\sim Multinomial([p_1, p_2, \ldots, p_k])$	$\hat{p_j} = \frac{n_j(\bar{x})}{N}$ where $n_j(\bar{x})$ is the count of category j in $\bar{x}$

Table 2: MLE estimated parameters for dataset $\bar{x} = [x_0, x_1, \cdots, x_{N-1}]$ where $N = length(\bar{x})$

3 Results of Naive Bayes Classifier

The MLE estimated parameters for all the features are summarized in Table 3 to Table 7. The F1 score values are summarized in Table 8.

References

[1] Aston Zhang et al. *Dive into Deep Learning*. 2021. URL: https://d2l.ai/.

Output class label y	Distribution	MLE estimated parameters
$y = 0$	$\sim Normal(\mu, \sigma^2)$	$\hat{\mu}_{X_1} = 2.02$, $\hat{\sigma}^2_{X_1} = 9.05$, $\hat{\mu}_{X_2} = 3.90$, $\hat{\sigma}^2_{X_2} = 78.42$
$y = 1$	$\sim Normal(\mu, \sigma^2)$	$\hat{\mu}_{X_1} = 0.02$, $\hat{\sigma}^2_{X_1} = 25.16$, $\hat{\mu}_{X_2} = 0.85$, $\hat{\sigma}^2_{X_2} = 230.03$
$y = 2$	$\sim Normal(\mu, \sigma^2)$	$\hat{\mu}_{X_1} = 8.02$, $\hat{\sigma}^2_{X_1} = 35.66$, $\hat{\mu}_{X_2} = -0.02$, $\hat{\sigma}^2_{X_2} = 4.00$

Table 3: MLE estimated parameters for feature X_1 and X_2

Output class label y	Distribution	MLE estimated parameters
$y = 0$	$\sim Bernoulli(p)$	$\hat{p}_{X_3} = 0.20$, $\hat{p}_{X_4} = 0.10$
$y = 1$	$\sim Bernoulli(p)$	$\hat{p}_{X_3} = 0.59$, $\hat{p}_{X_4} = 0.80$
$y = 2$	$\sim Bernoulli(p)$	$\hat{p}_{X_3} = 0.90$, $\hat{p}_{X_4} = 0.19$

Table 4: MLE estimated parameters for feature X_3 and X_4

Output class label y	Distribution	MLE estimated parameters
$y = 0$	$\sim Laplace(\mu, b)$	$\hat{\mu}_{X_5} = 0.07$, $\hat{b}_{X_5} = 1.98$, $\hat{\mu}_{X_6} = 0.87$, $\hat{b}_{X_6} = 5.97$
$y = 1$	$\sim Laplace(\mu, b)$	$\hat{\mu}_{X_5} = 0.38$, $\hat{b}_{X_5} = 0.99$, $\hat{\mu}_{X_6} = 0.35$, $\hat{b}_{X_6} = 5.99$
$y = 2$	$\sim Laplace(\mu, b)$	$\hat{\mu}_{X_5} = 0.79$, $\hat{b}_{X_5} = 3.00$, $\hat{\mu}_{X_6} = 0.21$, $\hat{b}_{X_6} = 3.06$

Table 5: MLE estimated parameters for feature X_5 and X_6

Output class label y	Distribution	MLE estimated parameters
$y = 0$	$\sim Exponential(\lambda)$	$\hat{\lambda}_{X_7} = 1.97$, $\hat{\lambda}_{X_8} = 3.93$
$y = 1$	$\sim Exponential(\lambda)$	$\hat{\lambda}_{X_7} = 2.98$, $\hat{\lambda}_{X_8} = 7.98$
$y = 2$	$\sim Exponential(\lambda)$	$\hat{\lambda}_{X_7} = 8.94$, $\hat{\lambda}_{X_8} = 14.68$

Table 6: MLE estimated parameters for feature X_7 and X_8

Output class label y	Distribution	MLE estimated parameters
$y = 0$	$\sim Multinomial([p_1, p_2, \ldots, p_k])$	$\hat{\hat{p}}_{X_9} = [0.2022, 0.2032, 0.2042, 0.1967, 0.1937]$, $\hat{\hat{p}}_{X_{10}} = [0.1213, 0.1236, 0.1257, 0.1277, 0.127, 0.1271, 0.1241, 0.1235]$
$y = 1$	$\sim Multinomial([p_1, p_2, \ldots, p_k])$	$\hat{\hat{p}}_{X_9} = [0.0977, 0.1984, 0.4047, 0.1583, 0.1409]$, $\hat{\hat{p}}_{X_{10}} = [0.1009, 0.0506, 0.0508, 0.1998, 0.1524, 0.1487, 0.2003, 0.0965]$
$y = 2$	$\sim Multinomial([p_1, p_2, \ldots, p_k])$	$\hat{\hat{p}}_{X_9} = [0.2052, 0.2997, 0.1029, 0.3417, 0.0505]$, $\hat{\hat{p}}_{X_{10}} = [0.1972, 0.0481, 0.0483, 0.1054, 0.1552, 0.153, 0.098, 0.1948]$

Table 7: MLE estimated parameters for feature X_9 and X_{10}

Output class label y	Training F1 score	Validation F1 score
$y = 0$	0.881	0.880
$y = 1$	0.878	0.878
$y = 2$	0.943	0.946

Table 8: F1 score for training and validation dataset