Bilkent University

CS 464 - Introduction to Machine Learning

Homework 01 Report

Emre Karataş
Section 01
22001641

November 12, 2023

Contents

1 Probability Review [10 pts]

Question 1.1 [4 pts]

To get two heads in a row, we should observe four different possible actions:

- Picking a blue coin from box one and getting two heads in a row

$$P(\text{Choose Box 1}) \times P(\text{Blue from Box 1}) \times P(\text{2 heads} \mid \text{Blue}) = \frac{1}{2} \times \frac{2}{3} \times \left(\frac{1}{2}\right)^2 = \frac{1}{12} \quad (1)$$

- Picking a yellow coin from box one and getting two heads in a row

$$P(\text{Choose Box 1}) \times P(\text{Yellow from Box 1}) \times P(\text{2 heads} \mid \text{Yellow}) = \frac{1}{2} \times \frac{1}{3} \times \left(\frac{1}{4}\right)^2 = \frac{1}{96} \quad (2)$$

- Picking a blue coin from box two and getting two heads in a row

$$P(\text{Choose Box 2}) \times P(\text{Blue from Box 2}) \times P(\text{2 heads} \mid \text{Blue}) = \frac{1}{2} \times \frac{1}{2} \times \left(\frac{1}{2}\right)^2 = \frac{1}{16} \quad (3)$$

- Picking a red coin from box two and getting two heads in a row

$$P(\text{Choose Box 2}) \times P(\text{Red from Box 2}) \times P(\text{2 heads} \mid \text{Red}) = \frac{1}{2} \times \frac{1}{2} \times \left(\frac{1}{10}\right)^2 = \frac{1}{400} \quad (4)$$

Adding all probabilities, we get:

$$\text{Total Probability} = \frac{1}{12} + \frac{1}{96} + \frac{1}{16} + \frac{1}{400} = \frac{127}{800} \approx 0.15875 \quad (5)$$

Question 1.2 [4 pts]

We can apply Bayes' Theorem to solve this question. Events in this question are:

- A: We picked a fair coin
- B: We got two heads in a row

We need to find $P(A \mid B)$, i.e., the probability that we picked a fair coin given that we got two heads in a row. Using Bayes' Theorem:

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

$P(B \mid A)$ is the conditional probability of getting two heads in a row if we pick a fair coin. That is, the equations (1) and (3) from Question 1.1 will give an answer for $P(B \mid A)$ since blue coins are fair in this setup. Therefore:

$$P(B|A) = \frac{1}{12} + \frac{1}{16} \quad (6)$$

$$= \frac{7}{48} \quad (7)$$

$P(A)$ is picking a fair coin from the system.

$$\begin{aligned}P(A) &= P(\text{picking Box 1}) \times P(\text{picking blue from Box 1}) \\ &\quad + P(\text{picking Box 2}) \times P(\text{picking blue from Box 2}) \\ &= 0.5 \times \frac{2}{3} + 0.5 \times 0.5 \\ &= \frac{5}{12} \qquad (8)\end{aligned}$$

We know $P(B)$ from Question 1.1. So, overall, we get:

$$P(A|B) = \frac{\frac{7}{48} \times \frac{5}{12}}{\frac{127}{800}} \qquad (9)$$

$$= \frac{3500}{5969} \approx 0.58636 \qquad (10)$$

Question 1.3 [2 pts]

Similarly, we can apply Bayes' Theorem to solve this question. Events in this question are:

- A: We picked a red coin
- B: We got two heads in a row

Using Bayes' Theorem:

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

$P(B \mid A)$ is the conditional probability of getting two heads in a row if we pick a red coin.

$$\left(\frac{1}{10}\right)^2 = \frac{1}{100} \qquad (11)$$

$P(A)$ is the probability of picking a red coin:

$$P(\text{Choose Box 2}) \times P(\text{Red from Box 2}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \qquad (12)$$

$P(B)$ is the total probability of getting 2 heads in a row, calculated in Question 1.1 as:

$$P(B) = \frac{127}{800}$$

So, overall:

$$P(A|B) = \frac{\frac{1}{100} \times \frac{1}{4}}{\frac{127}{800}} \qquad (13)$$

$$= \frac{2}{127} \approx 0.01574 \qquad (14)$$

2 MLE and MAP [20 pts]

Question 2.1 [8 pts]

Given a set of data points $x_1, x_2, \ldots, x_n$ that are normally distributed, we want to find the maximum likelihood estimator (MLE) for the mean μ.

The probability density function for a normal distribution is given by:

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

The likelihood function $L(\mu)$ for the sample $x_1, x_2, \ldots, x_n$ is:

$$L(\mu) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2)$$

Taking the logarithm of the likelihood function to simplify the multiplication into a sum (log-likelihood $l(\mu)$), we have:

$$l(\mu) = \sum_{i=1}^{n} \ln f(x_i; \mu, \sigma^2)$$

$$l(\mu) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)$$

$$l(\mu) = \sum_{i=1}^{n} \left[-\ln(\sigma\sqrt{2\pi}) - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$l(\mu) = -n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

To find the MLE for μ, we take the derivative of $l(\mu)$ with respect to μ, set it to zero, and solve for μ:

$$\frac{d}{d\mu}l(\mu) = \frac{d}{d\mu}\left(-n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$0 = \frac{d}{d\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$0 = \sum_{i=1}^{n}(x_i-\mu)$$

$$0 = \sum_{i=1}^{n} x_i - n\mu$$

$$n\mu = \sum_{i=1}^{n} x_i$$

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The MLE for μ is the sample mean.

Question 2.2 [8 pts]

The likelihood for the normal distribution is given by the product of the probability density function for each data point:

$$L(\mu|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

The prior distribution for μ, given it's an exponential distribution, is:

$$P(\mu) = \lambda e^{-\lambda\mu}$$

The posterior is proportional to the likelihood times the prior:

$$P(\mu|x_1, x_2, \ldots, x_n) \propto L(\mu|x_1, x_2, \ldots, x_n) \cdot P(\mu)$$

We can calculate the MAP estimate by taking the log of the posterior, which is the sum of the log-likelihood and the log prior, and then finding the derivative with respect to μ and setting it to zero to find the maximum.

The log of the posterior (up to a constant) is:

$$\log(P(\mu|x_1, x_2, \ldots, x_n)) = \sum_{i=1}^{n} \left[-\log(\sigma\sqrt{2\pi}) - \frac{(x_i - \mu)^2}{2\sigma^2} \right] - \lambda\mu$$

Taking the derivative with respect to μ, we get:

$$\frac{d}{d\mu} \log(P(\mu|x_1, x_2, \ldots, x_n)) = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} - \lambda$$

Setting this derivative to zero gives us the equation to solve for μ:

$$\sum_{i=1}^{n} \frac{x_i}{\sigma^2} - \frac{n\mu}{\sigma^2} - \lambda = 0$$

$$\Rightarrow \mu = \frac{\sum_{i=1}^{n} x_i - \lambda\sigma^2}{n}$$

Question 2.3 [4 pts]

Given that the normal distribution represents the dataset well with $\mu = 1$ and $\sigma = 1$, and a new data point x_{n+1} is found, we are to find the likelihood of $x_{n+1} = 2$ according to the probability density function (pdf) of the normal distribution.

The pdf of the normal distribution for a given x is:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Substituting $\mu = 1$ and $\sigma = 1$ into the pdf, we get:

$$f(2; 1, 1^2) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{(2-1)^2}{2\cdot 1^2}}$$

$$f(2; 1, 1^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}}$$

Upon calculating the above expression, the likelihood of observing $x_{n+1} = 2$ is approximately 0.242.

3 BBC News Classification [70 pts]

In this part of the homework, we are expected to examine given data files (x train.csv, y train.csv, x test.csv, y test.csv) and answer the asked questions.

Question 3.1 [10 pts]

These are four different parts for Question 3.1.

Question 3.1.1 [2.5 pts]

Percentages of each category in train and test of Y are as shown:

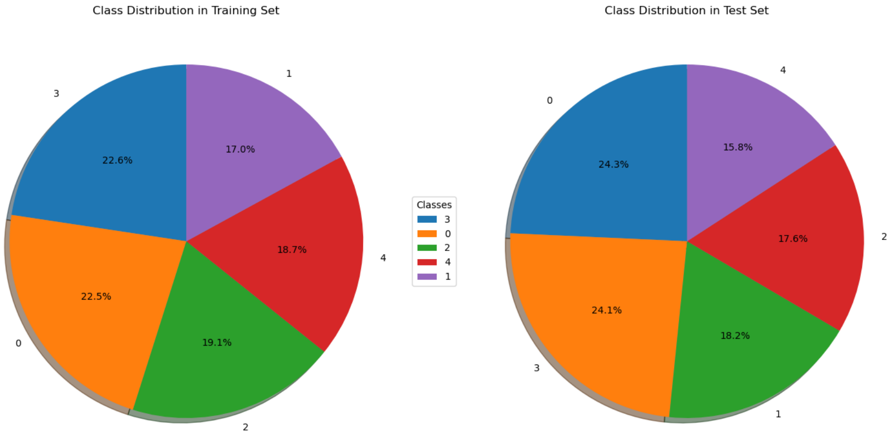


Figure 1: Class distribution in the training and test sets.

Class Distribution Counts

Category	Count
Class 0	377
Class 1	375
Class 2	319
Class 3	312
Class 4	284

Table 1: Category counts in the training set.

Category	Count
Class 0	135
Class 1	134
Class 2	101
Class 3	98
Class 4	88

Table 2: Category counts in the test set.

Question 3.1.2 [2.5 pts]

The prior probabilities of each class in the training set are given by the proportion of instances of each class relative to the total number of instances. The calculated prior probabilities are as follows:

Category	Prior Probability
Class 0	0.22495
Class 1	0.17036
Class 2	0.19136
Class 3	0.22615
Class 4	0.18716

Question 3.1.3 [2.5 pts]

The training set shows a slight imbalance in class distribution, with Class 0 being the most represented and Class 4 the least. This slight imbalance indicates that the dataset is not perfectly balanced but not heavily skewed towards any class. While this may not significantly impact the model’s performance, it is essential to be aware that an imbalanced training set can lead to a model bias towards more frequent classes, potentially affecting the model’s ability to generalize well to unseen data, especially for the minority classes.

Question 3.1.4 [2.5 pts]

The word ”alien” appears `3` times and the word ”thunder” appears `0` times in the training documents with the label ”Tech”. The log ratio of their occurrences within those documents are calculated as follows:

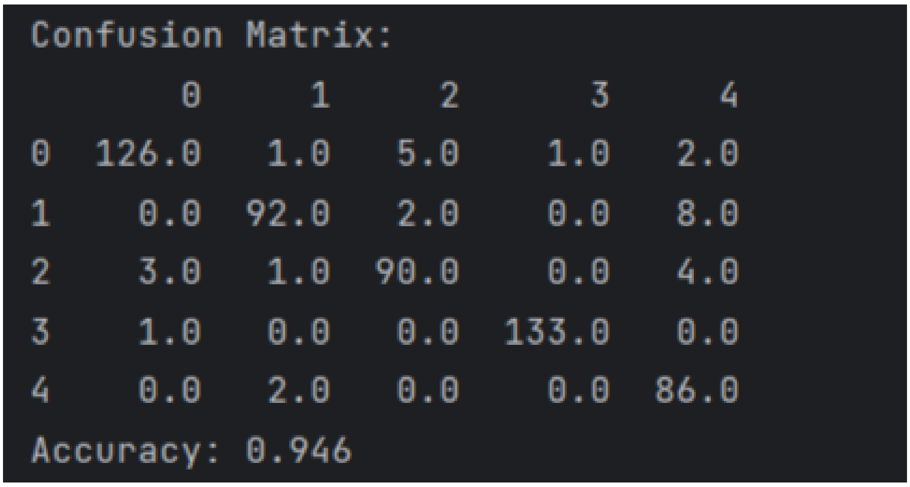
$$\ln(P(\text{alien}|Y = \text{Tech})) = \texttt{-9.999570161213173} \tag{15}$$

$$\ln(P(\text{thunder}|Y = \text{Tech})) = \texttt{-11.385864522333064} \tag{16}$$

Question 3.2 (Coding*) [20 pts]

Here are the accuracy and confusion matrix the results for the Multinomial Naive Bayes classifier.

- Accuracy: 0.946
- Confusion Matrix: See Figure 2.



```
Confusion Matrix:
       0     1     2      3     4
0  126.0   1.0   5.0    1.0   2.0
1    0.0  92.0   2.0    0.0   8.0
2    3.0   1.0  90.0    0.0   4.0
3    1.0   0.0   0.0  133.0   0.0
4    0.0   2.0   0.0    0.0  86.0
Accuracy: 0.946
```

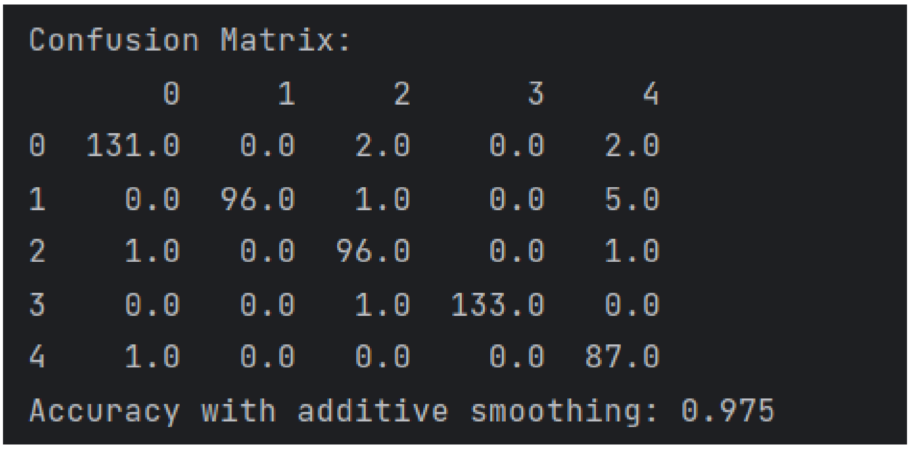
Figure 2: Confusion matrix and accuracy for the Multinomial Naive Bayes classifier.

Note:It is worth to note that while running this code, I got division by zero in log ”warning”, which did not interrupt running of the code but gave an warning in the Conda environment.

Question 3.3 (Coding*) [20 pts]

Here are the accuracy and confusion matrix the results for the Multinomial Naive Bayes classifier with Dirichet prior .

- Accuracy: 0.975
- Confusion Matrix: See Figure 3.



```
Confusion Matrix:
        0     1     2      3     4
0  131.0   0.0   2.0    0.0   2.0
1    0.0  96.0   1.0    0.0   5.0
2    1.0   0.0  96.0    0.0   1.0
3    0.0   0.0   1.0  133.0   0.0
4    1.0   0.0   0.0    0.0  87.0
Accuracy with additive smoothing: 0.975
```

Figure 3: Confusion matrix and accuracy for the Multinomial Naive Bayes classifier with Dirichet prior.

In our Naive Bayes classification, the introduction of the Dirichlet prior, or smoothing factor α, proved to be a significant factor for model fitting. By setting α to 1, we effectively accounted for the potential occurrence of every word in our dataset, avoiding the downsides of zero probabilities. This adjustment not only prevented overfitting but also improved the model's ability to handle new, unseen data. Reflecting on our results, the use of the Dirichlet prior led to an better results in accuracy, confirming its beneficial impact.

Question 3.4 (Coding*) [20 pts]

During the process of training the Bernoulli Naive Bayes classifier for task 3.4, an error was encountered, which halted the execution of the script. The error message was as follows:

```
Traceback (most recent call last):
  File "C:\Users\user\Desktop\CS464_HW01\q3main.py", line 243, in
      <module>
    main()
  File "C:\Users\user\Desktop\CS464_HW01\q3main.py", line 202, in
      main
    word_counts = X_train_binary.groupby(y_train).apply(lambda x:
        x.sum() + alpha)
  File "C:\Users\user\anaconda3\lib\site-packages\pandas\core\
     frame.py", line 8402, in groupby
    return DataFrameGroupBy(
  File "C:\Users\user\anaconda3\lib\site-packages\pandas\core\
     groupby\groupby.py", line 965, in __init__
    grouper, exclusions, obj = get_grouper(
  File "C:\Users\user\anaconda3\lib\site-packages\pandas\core\
     groupby\grouper.py", line 899, in get_grouper
    Grouping(
  File "C:\Users\user\anaconda3\lib\site-packages\pandas\core\
     groupby\grouper.py", line 542, in __init__
    raise ValueError(f"Grouper␣for␣’{t}’␣not␣1-dimensional")
ValueError: Grouper for ’<class␣’pandas.core.frame.DataFrame’>’
   not 1-dimensional
```

Error Analysis

During the implementation of section 3.4, we encountered an error related to the data dimensions being passed to the `groupby` method of a pandas DataFrame. This method requires a one-dimensional key to group the DataFrame, but it appears a two-dimensional DataFrame was provided instead, leading to a ValueError. To address this, we need to ensure that the key used for grouping is one-dimensional and aligns with the DataFrame's rows, representing the class labels for each observation.

Impact of the Error

As a direct consequence of this error, the script could not complete, and thus, the accuracy and confusion matrix for the Bernoulli Naive Bayes model could not be computed. For detailed code and the specific issues encountered, please refer to the code listing in section 3.4.

Comparison of Naive Bayes Models

Despite the challenges in implementation, it is important to discuss the theoretical differences between the Bernoulli and Multinomial Naive Bayes models. The Bernoulli model, which considers the presence or absence of features, is contrasted with the Multinomial model, which accounts for feature frequency. Theoretically, the Bernoulli model should be more suitable for binary or boolean features, while the Multinomial model is expected to excel in situations where the frequency of feature occurrence provides significant information for classification.

Theoretical Implications

Although the error prevented the empirical comparison of models, we anticipate that the Bernoulli model's accuracy would reflect its appropriateness for the given dataset, as it simplifies the feature representation to binary terms. This is particularly useful in datasets where feature presence carries more weight than feature count. On the other hand, the Multinomial model's reliance on word frequency makes it a strong candidate for text classification tasks where the context provided by word count is essential for making accurate predictions.

Conclusions

In conclusion, the choice between a Bernoulli or a Multinomial Naive Bayes model should be informed by the dataset's characteristics and the relative importance of feature presence versus feature frequency. Future work will involve correcting the implementation issues to provide empirical support for these theoretical considerations.