Text Classification Using Naive Bayes

This document's content primarily covers *text classification* with a focus on the Naive Bayes method and its application to categorizing text, such as documents or emails, into predefined categories. I'll break down this chapter in depth and provide illustrations as requested, starting with the following four sections:

1. **Introduction to Text Classification and Naive Bayes Theory**
2. **Naive Bayes Multinomial Model in Text Classification**
3. **Naive Bayes Bernoulli Model in Text Classification**
4. **Feature Selection for Text Classification**

Here's the detailed breakdown of the first part:

Part 1: Introduction to Text Classification and Naive Bayes Theory

Overview of Text Classification

Text classification is the process of assigning a predefined category to a given document based on its content. This task arises frequently in information retrieval applications, such as:

- Email sorting and spam filtering.
- Sentiment analysis in product reviews.
- Identifying content topics in news articles.

Classification tasks vary in complexity:

- **Binary classification**: Assigning one of two classes (e.g., spam vs. not-spam).
- **Multiclass classification**: Assigning one of several classes (e.g., classifying news articles into sports, technology, finance, etc.).

Key Concepts in Classification

- **Classes**: Defined categories to which documents are assigned (e.g., politics, sports).
- **Document Space (X)**: A high-dimensional space where each document is represented by its attributes (e.g., word frequency).
- **Training Set (D)**: A collection of labeled documents used to train the classifier.
- **Classifier (γ)**: A function that assigns classes to new, unseen documents.

The goal of classification is to create a function that assigns the correct class to new documents, achieving high accuracy on a test set.

Types of Classification Problems

1. **One-of Problems**: A document can only belong to one class (e.g., news article topic).
2. **Any-of Problems**: A document may belong to multiple classes (e.g., a document discussing the Olympics could belong to both "sports" and "China").

Naive Bayes Classifier Theory

Naive Bayes (NB) is a popular method for classification, leveraging probability and assuming that each feature contributes independently to the classification decision. Despite this "naive" independence assumption, NB often performs well in practice.

Supervised Learning in Naive Bayes

Naive Bayes relies on **supervised learning**, where the model learns from labeled training data. The main task here is to find a classification function that maximizes the probability of assigning the correct class to new documents.

The process involves:

1. Calculating **prior probabilities** for each class based on training data.
2. Calculating the **likelihood** of terms in each document given a class.
3. Applying Bayes' theorem to assign a document to the class with the highest posterior probability.

Numerical Example of Classifying Documents with Naive Bayes

Problem Setup

Consider a training set with documents labeled by topic. For simplicity, suppose we have documents about two topics: *China* and *Japan*. A document contains various words (tokens), and we aim to classify it based on word frequency.

Training Data Example

Let's say we have a dataset where each document contains terms with a label indicating the topic:

Document ID	Words in Document	Topic
1	Chinese Beijing Chinese	China
2	Chinese Chinese Shanghai	China
3	Chinese Macao	China
4	Tokyo Japan Chinese	Japan

Using this data, we'll calculate:

1. **Prior Probabilities**: The likelihood of each class based on document counts.
2. **Conditional Probabilities**: The probability of each word given the class.

Calculating Priors and Likelihoods

- **Prior for China ($P(\textbf{China})$)**: 3 out of 4 documents are about China, so $P(\text{China}) = \frac{3}{4}$.
- **Prior for Japan ($P(\textbf{Japan})$)**: 1 out of 4 documents, so $P(\text{Japan}) = \frac{1}{4}$.

For the word "Chinese":

- **Likelihood $P(\textbf{Chinese}|\textbf{China})$**: There are 5 instances of "Chinese" in China documents out of 8 words total for China, so $P(\text{Chinese}|\text{China}) = \frac{5}{8}$.
- **Likelihood $P(\textbf{Chinese}|\textbf{Japan})$**: 1 instance in Japan documents out of 3 words, so $P(\text{Chinese}|\text{Japan}) = \frac{1}{3}$.

Classifying a New Document

For a new document, "Chinese Chinese Tokyo Japan," we use the probabilities calculated to assign it to the most likely topic class.

Additional Exercise

1. Calculate the posterior probability of the new document belonging to China or Japan.
2. Evaluate how adding new words (e.g., Beijing or Macao) affects classification.

1.4 Naive Bayes Formula and Interpretation

The NB classifier calculates the probability of a class c given a document d using:

$$P(c \mid d) \propto P(c) \prod_{k=1}^{n_d} P\left(t_k \mid c\right)$$

Part 2: Naive Bayes Multinomial Model in Text Classification

The Multinomial Naive Bayes (NB) model is one of the simplest and most effective methods for text classification. This model is often used when working with text data and is particularly well-suited for applications where the goal is to classify documents based on word frequency.

Multinomial Naive Bayes Model Overview

In the Multinomial NB model:

- **Word Occurrence**: Each word's occurrence within a document contributes to the probability of the document being in a particular class.
- **Feature Independence**: The model assumes independence among words, meaning that the presence of one word does not affect the probability of another word in the same document, given the class.

For classification, the probability of a document d belonging to class c is calculated as:

$$P(c|d) \propto P(c) \prod_{k=1}^{n_d} P(t_k|c)$$

where:

- $P(c)$: Prior probability of class c.
- $P(t_k|c)$: Probability of term t_k occurring in class c.
- n_d: Total number of words in document d.

To find the best class for a document, the model selects the class with the highest **posterior probability** $P(c|d)$, also known as the **Maximum a Posteriori (MAP)** estimation.

Computing Probabilities with the Multinomial Model

1. Prior Probability Calculation

For each class c, the prior probability $P(c)$ is calculated based on the fraction of documents in the training set belonging to that class:

$$P(c) = \frac{\text{Number of documents in class } c}{\text{Total number of documents}}$$

2. Conditional Probability Calculation

Conditional probabilities $P(t|c)$ represent the likelihood of each term t given class c, calculated as:

$$P(t|c) = \frac{\text{Count of term } t \text{ in documents of class } c + 1}{\text{Total terms in class } c + B}$$

where B is the number of unique terms in the vocabulary, and Laplace smoothing is applied by adding 1 to both the numerator and denominator.

Avoiding Floating Point Underflow

In practical applications, multiplying many probabilities can lead to very small numbers, resulting in **floating point underflow**. To address this, we compute the **logarithm** of probabilities, allowing us to sum logs rather than multiply probabilities:

$$\text{MAP class} = \arg\max_{c \in C} \left[\log P(c) + \sum_{k=1}^{n_d} \log P(t_k|c) \right]$$

This approach provides numerical stability and makes calculations feasible for longer documents.

Numerical Illustration: Classifying with the Multinomial Naive Bayes Model

Consider the training dataset:

Document ID	Words in Document	Class
1	Chinese Beijing Chinese	China
2	Chinese Chinese Shanghai	China
3	Chinese Macao	China
4	Tokyo Japan Chinese	Japan

To classify the test document: "Chinese Chinese Chinese Tokyo Japan," we proceed as follows:

Step 1: Calculate Priors

$$P(\text{China}) = \frac{3}{4} = 0.75 \quad \text{and} \quad P(\text{Japan}) = \frac{1}{4} = 0.25$$

Step 2: Calculate Conditional Probabilities with Laplace Smoothing

For the term "Chinese":

$$P(\text{Chinese}|\text{China}) = \frac{5+1}{8+6} = \frac{6}{14} = 0.4286$$

$$P(\text{Chinese}|\text{Japan}) = \frac{1+1}{3+6} = \frac{2}{9} = 0.2222$$

For the term "Tokyo":

$$P(\text{Tokyo}|\text{China}) = \frac{0+1}{8+6} = \frac{1}{14} = 0.0714$$

$$P(\text{Tokyo}|\text{Japan}) = \frac{1+1}{3+6} = \frac{2}{9} = 0.2222$$

For the term "Japan":

$$P(\text{Japan}|\text{China}) = \frac{0+1}{8+6} = 0.0714$$

$$P(\text{Japan}|\text{Japan}) = \frac{1+1}{3+6} = 0.2222$$

Step 3: Compute the Log Probabilities for Each Class

For the "China" class:

$$\log P(\text{China}) + 3 \times \log(0.4286) + \log(0.0714) + \log(0.0714) \approx -0.2877 - 2.5595 - 5.2877 = -8.1349$$

For the "Japan" class:

$$\log P(\text{Japan}) + 3 \times \log(0.2222) + \log(0.2222) + \log(0.2222) \approx -1.3863 - 4.8503 - 2.7726 = -9.0092$$

Step 4: Classification Decision

Since $-8.1349 > -9.0092$, the document is classified as **China**.

This example demonstrates the core process of text classification with the Multinomial NB model, highlighting its simplicity and effectiveness in practical applications.

Part 3: Naive Bayes Bernoulli Model in Text Classification

The Bernoulli Naive Bayes (NB) model is another popular variation of the Naive Bayes classifier, differing from the Multinomial model in how it handles text data. While the Multinomial NB model accounts for the frequency of words, the Bernoulli NB model focuses on the presence or absence of words in a document. This makes the Bernoulli model particularly suitable for binary features and is often applied in tasks where a binary document representation is more appropriate.

Key Characteristics of the Bernoulli Model

In the Bernoulli NB model:

- **Binary Representation**: Each term is treated as a binary feature—either present (1) or absent (0) in the document.
- **Conditional Probabilities**: For each term in the vocabulary, the model calculates the probability that it will appear in a document given a specific class.
- **Absence Information**: Unlike the Multinomial model, the Bernoulli model incorporates the probability of terms not appearing in the document, which can influence the classification decision.

This model is especially useful for short documents, where individual terms' presence or absence can be strong indicators of class membership. However, it may struggle with long documents where term frequency would be a more informative feature.

Bernoulli Model Formula

The probability of a document d belonging to class c in the Bernoulli model is:

$$P(c|d) \propto P(c) \prod_{t \in V_d} P(t|c) \prod_{t \notin V_d} (1 - P(t|c))$$

where:

- V_d is the set of terms present in document d.
- $P(t|c)$ is the probability of term t appearing in documents of class c.
- $(1 - P(t|c))$ accounts for terms that do not appear in the document but still affect the classification.

Steps for Calculating Probabilities in the Bernoulli Model

1. **Prior Probability Calculation**:

$$P(c) = \frac{\text{Number of documents in class } c}{\text{Total number of documents}}$$

2. **Conditional Probability for Term Presence and Absence**:

$$P(t|c) = \frac{\text{Number of documents in class } c \text{ containing term } t + 1}{\text{Number of documents in class } c + 2}$$

3. **Classification Decision**:

$$\text{MAP class} = \arg\max_{c \in C} \left[\log P(c) + \sum_{t \in V_d} \log P(t|c) + \sum_{t \notin V_d} \log(1 - P(t|c)) \right]$$

Numerical Example: Classifying with the Bernoulli Naive Bayes Model

Consider the same dataset as before, where the training set is as follows:

Document ID	Words in Document	Class
1	Chinese Beijing Chinese	China
2	Chinese Chinese Shanghai	China
3	Chinese Macao	China
4	Tokyo Japan Chinese	Japan

We want to classify a new document: *Chinese Chinese Tokyo Japan.*

1. **Step 1: Calculate Priors**

$$P(\text{China}) = \frac{3}{4} = 0.75, \quad P(\text{Japan}) = \frac{1}{4} = 0.25$$

2. **Step 2: Calculate Conditional Probabilities with Laplace Smoothing**

   - For the term "Chinese":

$$P(\text{Chinese}|\text{China}) = \frac{3+1}{3+2} = 0.8, \quad P(\text{Chinese}|\text{Japan}) = \frac{1+1}{1+2} = 0.67$$

   - For the term "Tokyo":

$$P(\text{Tokyo}|\text{China}) = \frac{0+1}{3+2} = 0.2, \quad P(\text{Tokyo}|\text{Japan}) = \frac{1+1}{1+2} = 0.67$$

   - For the term "Japan":

$$P(\text{Japan}|\text{China}) = 0.2, \quad P(\text{Japan}|\text{Japan}) = 0.67$$

3. **Step 3: Compute the Log-Probabilities for Each Class**

$$\log P(\text{China}) + 2 \times \log(0.8) + \log(0.2) + \log(0.2) = -4.059$$

$$\log P(\text{Japan}) + 2 \times \log(0.67) + \log(0.67) + \log(0.67) = -3.405$$

4. **Step 4: Classification Decision** Since $-3.405 > -4.059$, the document is classified as **Japan**.

This example highlights how the Bernoulli model accounts for the presence and absence of terms, which can lead to different classifications compared to the Multinomial model.

Part 4: Feature Selection in Text Classification

Feature selection is a critical process in text classification, aiming to reduce the vocabulary size by selecting the most relevant terms. This not only improves the efficiency of the Naive Bayes classifiers but also often enhances classification accuracy by removing irrelevant or noisy features. In this section, we cover the main approaches to feature selection in text classification and examine their impact on the Multinomial and Bernoulli Naive Bayes models.

Objectives of Feature Selection

- **Efficiency**: Reducing the number of features makes the training and classification process faster.
- **Improved Accuracy**: By eliminating noise (irrelevant or misleading features), the classifier can achieve better accuracy on new, unseen data.
- **Robustness**: Focusing on core features can make the model less sensitive to variations in the data, such as changes in word usage patterns.

Key Feature Selection Techniques

1. **Mutual Information**
2. **Chi-Square Test (χ^2)**
3. **Frequency-Based Selection**

Each of these techniques computes a "score" for each feature based on different criteria. The top k features with the highest scores are then selected for use in the classification model.

1. Mutual Information (MI)

Mutual Information (MI) measures how much knowing the presence or absence of a term in a document informs the correct classification. MI captures the association between each term and each class, favoring terms that are strong indicators of class membership.

The formula for MI is:

$$I(U;C) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(U = e_t, C = e_c) \log \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}$$

where:

- U represents whether a term appears (1) or does not appear (0) in a document.
- C represents whether a document belongs to the class c (1) or not (0).

2. Chi-Square Test (χ^2)

The Chi-Square test (χ^2) assesses the independence between a term and the class label. If a term's presence is strongly associated with a particular class, it is considered a valuable feature. The χ^2 test is effective at identifying terms that are good discriminators between classes, especially for rare terms.

The formula for χ^2 is:

$$\chi^2(t,c) = \frac{(N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01})(N_{11} + N_{10})(N_{10} + N_{00})(N_{01} + N_{00})}$$

where:

- N_{11} is the number of documents in class c that contain term t.
- N_{10} is the number of documents not in class c that contain term t.
- N_{01} is the number of documents in class c that do not contain term t.
- N_{00} is the number of documents not in class c that do not contain term t.

3. Frequency-Based Feature Selection

Frequency-based selection is simpler than MI and χ^2 as it relies on the raw frequency of terms in the class:

- **Document Frequency**: Number of documents in class c containing the term t.
- **Collection Frequency**: Total occurrences of term t in documents of class c.

This method is particularly effective when selecting a large number of features. While it may include some non-informative terms, it often performs well in practice when used in conjunction with the Multinomial NB model, which can handle a larger number of terms efficiently.

Comparison of Feature Selection Techniques

Each of these methods has unique characteristics:

- **Mutual Information (MI)**: Best for terms that strongly correlate with specific classes.
- **Chi-Square (χ^2)**: Ideal for discriminative terms, especially for small, significant terms.
- **Frequency-Based Selection**: A good choice when a large number of features is needed, although it may include common terms with less class specificity.

Impact on Model Performance

Feature selection's effect on the Multinomial and Bernoulli models can be summarized as follows:

- **Multinomial NB**: Benefits from a larger vocabulary as it considers term frequency.
- **Bernoulli NB**: Prefers a smaller, more focused vocabulary since it only considers term presence or absence.

Selecting an optimal feature set often involves trial and error. In practice, feature selection is iteratively refined to achieve the best balance between efficiency and classification accuracy.

Numerical Illustration of Feature Selection

Suppose we have the following frequency-based data for terms related to different classes (China, UK, and poultry):

Term	**Class: China**	**Class: UK**	**Class: Poultry**
Chinese	High	Low	Low
Beijing	High	Low	Low
Tokyo	Low	Low	Low
Poultry	Low	Low	High
Chicken	Low	Low	High

Table 1: Frequency-Based Feature Selection for Various Classes

Using feature selection based on MI, we observe:

- High MI for "Chinese" and "Beijing" with China.
- High MI for "Poultry" and "Chicken" with the poultry class.

By selecting the top features based on MI or χ^2, we create a more efficient model. By testing with varying feature set sizes, we can find an optimal number for the specific text classification task.