Multinomial Naïve Bayes for Text Classification

20221061

Yoon seo Oh

1. Intro to the theroy

Bayes' theorem is utilized to calculate conditional probabilities and finds extensive application in solving classification problems within machine learning.

Bayes' theorem is represented by the following equation: $P(A|B)=P(B)P(B|A)\cdot P(A)$

Here:

- $P(A|B)$: Conditional probability of A given B (posterior probability)
- $P(B|A)$: Conditional probability of B given A (likelihood)
- $P(A)$: Prior probability of A
- $P(B)$: Probability of B

The Naïve Bayes algorithm is based on Bayes' theorem and is particularly used for classification problems. It calculates the probability that a given set of input features belongs to each class (label) and predicts the class with the highest probability.

The equation $P(L|features)=P(features)P(features|L)\cdot P(L)$ is used in the Naïve Bayes classifier.

Here:

- $P(L|features)$: Posterior probability of each label (Li) given observed features
- $P(features|L)$: Likelihood of features given a label (Li)
- $P(L)$: Prior probability of each label (Li)
- $P(features)$: Probability of observed features

Using this formula, the Naïve Bayes classifier computes the probability that each class (label) belongs to given features, enabling the prediction of the class with the highest probability. The Multinomial Naïve Bayes model employs distributions that describe the probabilities of occurrence for different categories in discrete situations, thereby modeling the occurrence frequencies for categorical variables.

2. My implementation

- **Training Phase**:
  - The code initializes dictionaries **category_word_probs** and **category_counts** to store word probabilities and word counts for each category, respectively.
  - It iterates through the training data and calculates word probabilities and counts for each category using Laplace smoothing.
- **Testing Phase**:
  - It iterates through the test data and predicts the category for each document.
  - For each document, it calculates the log probability of the document belonging to each category using the multinomial naive Bayes formula.
  - The predicted category is chosen based on the highest calculated log probability for each document.
  - The accuracy of the predictions is calculated using **accuracy_score** by comparing predicted categories with the actual categories in the test set.

3. Performance evaluation result using test dataset

- Insufficient Consideration of Class Imbalance: Without considering class imbalance in the data, comparing performance across different classes might be challenging.
- Balancing Sample Distribution: If class imbalance exists, considering preprocessing techniques like sampling methods to balance classes might be beneficial.