Naive Bayes Classifier Theory

1 Introduction

The Naive Bayes classifier is a probabilistic machine learning algorithm based on Bayes' theorem with the assumption of independence between features. It is commonly used for classification tasks and is particularly effective when the dimensionality of the data is high.

2 Bayes' Theorem

The Naive Bayes classifier is grounded in Bayes' theorem, which relates the conditional and marginal probabilities of random variables. For a given class label y and feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, Bayes' theorem states:

$$P(y \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid y) \cdot P(y)}{P(\mathbf{x})} \tag{1}$$

where: - $P(y \mid \mathbf{x})$ is the posterior probability of class y given features $\mathbf{x}$. - $P(\mathbf{x} \mid y)$ is the likelihood of features $\mathbf{x}$ given class y. - $P(y)$ is the prior probability of class y. - $P(\mathbf{x})$ is the marginal probability of features $\mathbf{x}$.

3 Naive Assumption

The Naive Bayes classifier assumes that features are conditionally independent given the class label. This simplifies the likelihood calculation:

$$P(\mathbf{x} \mid y) = \prod_{j=1}^{d} P(x_j \mid y) \tag{2}$$

Thus, the posterior probability can be computed as:

$$P(y \mid \mathbf{x}) = \frac{P(y) \prod_{j=1}^{d} P(x_j \mid y)}{P(\mathbf{x})} \tag{3}$$

Since $P(\mathbf{x})$ is constant for all classes, it can be ignored in classification. Thus, we classify a sample $\mathbf{x}$ by finding the class y that maximizes the posterior probability:

$$\hat{y} = \arg\max_{y} \left(P(y) \prod_{j=1}^{d} P(x_j \mid y) \right) \tag{4}$$

4 Multinomial Naive Bayes

In the Multinomial Naive Bayes model, the features are assumed to be counts or frequencies. The likelihood is modeled using a multinomial distribution:

$$P(\mathbf{x} \mid y) = \frac{\prod_{j=1}^{d} (P(x_j \mid y))^{x_j}}{\text{Normalization Factor}} \tag{5}$$

where x_j represents the count or frequency of feature j in the sample.

5 Laplace Smoothing

To handle zero probabilities for unseen features, Laplace smoothing (additive smoothing) is applied:

$$P(x_j \mid y) = \frac{N_{xyj} + \alpha}{N_{xy} + \alpha \cdot V} \tag{6}$$

where: - N_{xyj} is the count of feature j given class y. - N_{xy} is the total count of all features for class y. - α is the smoothing parameter (often set to 1). - V is the number of possible values for the feature.