Custom Bayes Classifier: A Comprehensive Approach to Addressing Class Imbalance

Rashin Rahnamoun

Mathematical Formulation

Class Probability Adjustment

I begin by adjusting the class probabilities with an advanced approach that accounts for class imbalance using weighted priors.

$$P(y = C_k) = \frac{N_k + \alpha}{N + \alpha \cdot K}$$

where:

- N_k is the number of training instances in class C_k,
- N is the total number of training instances,
- K is the total number of classes,
- α is the smoothing parameter (often set to 1).

Feature Probability Adjustment

The feature probabilities are adjusted using a modified Laplace smoothing technique to better handle class imbalances.

$$P(x_i|y = C_k) = \frac{N_{ik} + \alpha}{N_k + \alpha \cdot V}$$

where:

- N_{ik} is the number of instances where feature x_i occurs in class C_k,
- N_k is the total number of instances in class C_k,
- V is the size of the vocabulary (number of distinct features),
- α is the Laplace smoothing parameter.

Theoretical Foundation

Lemma 1. *For any class C_k, the adjusted class probability satisfies the inequality:*

$$0 < \frac{N_k + \alpha}{N + \alpha \cdot K} < 1$$

Proof. To prove the inequality $0 < \frac{N_k+\alpha}{N+\alpha \cdot K} < 1$,I start by analyzing the numerator and denominator of the fraction separately.

First, consider the numerator $N_k + \alpha$. By definition, N_k represents the number of training instances in class C_k. Since N_k is a count of instances, it is always non-negative, i.e., $N_k \geq 0$. The parameter α is a smoothing parameter, often chosen to be a positive value (commonly 1), thus $\alpha > 0$.

Given these conditions, the numerator $N_k + \alpha$ is strictly positive:

$$N_k + \alpha > 0$$

Next, consider the denominator $N + \alpha \cdot K$. Here, N represents the total number of training instances across all classes, and K is the total number of classes. Since there are instances present in the training set, $N > 0$. Additionally, α is a positive constant and $K \geq 1$, so $\alpha \cdot K > 0$.

Adding these two positive quantities gives a strictly positive denominator:

$$N + \alpha \cdot K > 0$$

Since both the numerator and the denominator are strictly positive, the fraction $\frac{N_k+\alpha}{N+\alpha \cdot K}$ is also strictly positive:

$$\frac{N_k + \alpha}{N + \alpha \cdot K} > 0$$

To establish the upper bound,I note that the numerator $N_k + \alpha$ is less than the denominator $N + \alpha \cdot K$ because $N_k \leq N$ (as N_k is a part of the total N) and α is the same in both terms but multiplied by K in the denominator.

Therefore, the fraction is always less than 1:

$$\frac{N_k + \alpha}{N + \alpha \cdot K} < 1$$

Combining these results,I confirm that:

$$0 < \frac{N_k + \alpha}{N + \alpha \cdot K} < 1$$

Thus, the adjusted class probability satisfies the required inequality. □

Lemma 2. *The sum of adjusted class probabilities for all classes is equal to 1:*

$$\sum_{k=1}^{K} P(y = C_k) = 1$$

Proof. To show that the sum of the adjusted class probabilities for all classes equals 1, we start by expressing the sum:

$$\sum_{k=1}^{K} P(y = C_k)$$

Substituting the adjusted class probability formula $P(y = C_k) = \frac{N_k+\alpha}{N+\alpha \cdot K}$ into the sum, I get:

$$\sum_{k=1}^{K} \frac{N_k + \alpha}{N + \alpha \cdot K}$$

I can factor out the constant denominator $N + \alpha \cdot K$:

$$\frac{1}{N + \alpha \cdot K} \sum_{k=1}^{K} (N_k + \alpha)$$

Next, I need to simplify the sum inside the parentheses:

$$\sum_{k=1}^{K} (N_k + \alpha)$$

This sum can be separated into two parts:

$$\sum_{k=1}^{K} N_k + \sum_{k=1}^{K} \alpha$$

The first part, $\sum_{k=1}^{K} N_k$, is simply the total number of training instances N:

$$\sum_{k=1}^{K} N_k = N$$

The second part, $\sum_{k=1}^{K} \alpha$, is the sum of the smoothing parameter α repeated K times:

$$\sum_{k=1}^{K} \alpha = \alpha \cdot K$$

Combining these results,I get:

$$\sum_{k=1}^{K} (N_k + \alpha) = N + \alpha \cdot K$$

Substituting this back into our fraction,I have:

$$\frac{1}{N + \alpha \cdot K} (N + \alpha \cdot K)$$

Since the numerator and the denominator are the same, the fraction simplifies to 1:

$$\frac{N + \alpha \cdot K}{N + \alpha \cdot K} = 1$$

Therefore, I have shown that:

$$\sum_{k=1}^{K} P(y = C_k) = 1$$

Thus, the sum of the adjusted class probabilities for all classes equals 1. □

Theorem 1. *The adjusted feature probability for any feature x_i given class C_k is a valid probability:*

$$0 < \frac{N_{ik} + \alpha}{N_k + \alpha \cdot V} < 1$$

Proof. To establish that the adjusted feature probability is a valid probability, I must show that $0 < \frac{N_{ik}+\alpha}{N_k+\alpha \cdot V} < 1$.

First, consider the numerator $N_{ik} + \alpha$. By definition, N_{ik} represents the number of instances in which feature x_i occurs in class C_k. Since N_{ik} is a count, it is non-negative, i.e., $N_{ik} \geq 0$. The parameter α is a positive smoothing parameter, so $\alpha > 0$.

Therefore, the numerator $N_{ik} + \alpha$ is strictly positive:

$$N_{ik} + \alpha > 0$$

Next, consider the denominator $N_k + \alpha \cdot V$. Here, N_k represents the total number of instances in class C_k, which is always positive ($N_k > 0$), and $\alpha \cdot V$ is a positive term (with $\alpha > 0$ and $V > 0$).

Thus, the denominator $N_k + \alpha \cdot V$ is strictly positive:

$$N_k + \alpha \cdot V > 0$$

Since both the numerator and denominator are positive, the fraction $\frac{N_{ik}+\alpha}{N_k+\alpha \cdot V}$ is also positive:

$$\frac{N_{ik} + \alpha}{N_k + \alpha \cdot V} > 0$$

To establish the upper bound, note that $N_{ik} \leq N_k$ and hence:

$$N_{ik} + \alpha \leq N_k + \alpha$$

This implies:

$$\frac{N_{ik} + \alpha}{N_k + \alpha \cdot V} \leq \frac{N_k + \alpha}{N_k + \alpha \cdot V}$$

Since $\alpha \cdot V \geq \alpha$, the fraction:

$$\frac{N_k + \alpha}{N_k + \alpha \cdot V} < 1$$

Thus, the adjusted feature probability is always less than 1:

$$\frac{N_{ik} + \alpha}{N_k + \alpha \cdot V} < 1$$

Combining these results, I confirm that:

$$0 < \frac{N_{ik} + \alpha}{N_k + \alpha \cdot V} < 1$$

Thus, the adjusted feature probability is a valid probability. □

To handle the inherent uncertainty and variability in feature classification, we can employ a *probabilistic approach*. Let $P(x_i \mid C_k)$ represent the conditional probability of feature x_i belonging to class C_k. This probability can be expressed as:

$$P(x_i \mid C_k) = \frac{P(x_i \mid y = C_k)}{\sum_{l=1}^{K} P(x_i \mid y = C_l)}$$

Here, the probability of feature x_i being in class C_k is normalized by the sum of probabilities across all classes, ensuring it is properly scaled.

Additionally, a softmax function can be applied to adjust the probabilities based on feature relevance and class distributions. The adjusted probability $P'(x_i \mid C_k)$ is defined as:

$$P'(x_i \mid C_k) = \frac{e^{\alpha \cdot P(x_i|y=C_k)}}{\sum_{l=1}^{K} e^{\alpha \cdot P(x_i|y=C_l)}}$$

where α is a scaling parameter that influences the spread of the probability values, allowing for better handling of uncertainty in the classification process.

Differential Equation for Class Probability Adjustment

To incorporate fuzzy logic into the class probability adjustment, I can use a differential equation to model the dynamic adjustment of probabilities over time:

$$\frac{dP(y = C_k)}{dt} = \beta \left[\frac{N_k(t) + \alpha}{N(t) + \alpha \cdot K} - P(y = C_k) \right]$$

where β is a positive constant that controls the rate of adjustment. This differential equation models how the class probabilities evolve over time t based on the difference between the current probability and the adjusted probability.

To solve the differential equation, I follow these steps:

Let $P_{desired}(t) = \frac{N_k(t)+\alpha}{N(t)+\alpha \cdot K}$. Then the differential equation becomes:

$$\frac{dP(y = C_k)}{dt} = \beta \left(P_{desired}(t) - P(y = C_k) \right)$$

This is a first-order linear differential equation of the form:

$$\frac{dP}{dt} + \beta P = \beta P_{desired}(t)$$

To solve it, I use the integrating factor method.

1. The integrating factor $\mu(t)$ is:

$$\mu(t) = e^{\int \beta \, dt} = e^{\beta t}$$

2. Multiply both sides of the differential equation by the integrating factor:

$$e^{\beta t} \frac{dP(y = C_k)}{dt} + \beta e^{\beta t} P(y = C_k) = \beta e^{\beta t} P_{desired}(t)$$

3. The left side is the derivative of $e^{\beta t} P(y = C_k)$:

$$\frac{d}{dt} \left(e^{\beta t} P(y = C_k) \right) = \beta e^{\beta t} P_{desired}(t)$$

4. Integrate both sides with respect to t:

$$e^{\beta t} P(y = C_k) = \int \beta e^{\beta t} P_{desired}(t) \, dt + C$$

5. Solve for $P(y = C_k)$:

$$P(y = C_k) = e^{-\beta t} \left[\int \beta e^{\beta t} P_{desired}(t) \, dt + C \right]$$

6. Substitute $P_{desired}(t) = \frac{N_k(t)+\alpha}{N(t)+\alpha \cdot K}$:

$$P(y = C_k) = e^{-\beta t} \left[\int \beta e^{\beta t} \frac{N_k(t) + \alpha}{N(t) + \alpha \cdot K} \, dt + C \right]$$

To find the constant of integration C, use the initial condition $P(y = C_k)(0)$:

$$P(y = C_k)(0) = C$$

So, the solution to the differential equation is:

$$P(y = C_k)(t) = P(y = C_k)(0) e^{-\beta t} + \left(1 - e^{-\beta t}\right) \frac{N_k(t) + \alpha}{N(t) + \alpha \cdot K}$$

Step 3: Conclusion

The solution to the differential equation shows that the class probability $P(y = C_k)(t)$ evolves over time as a combination of the initial probability and the desired probability based on the current data. As $t \to \infty$, the term involving $e^{-\beta t}$ vanishes, and the class probability $P(y = C_k)$ approaches the desired probability $\frac{N_k(t)+\alpha}{N(t)+\alpha \cdot K}$.

Thus, the differential equation ensures that class probabilities converge to the empirical class probabilities, reflecting the balance between initial guesses and observed data trends. The learning rate β controls how quickly this convergence occurs. The results showed that my dataset achieved only a 1 percent improvement compared to the original Naive Bayes algorithm.