Naive Bayes Classifier

Naive Bayes is a probabilistic machine learning algorithm used for classification tasks, based on Bayes' Theorem with a naive assumption that features are independent of each other.

- **Bayes' Theorem:**

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

In classification terms:

$$P(\text{Class}|\text{Features}) = \frac{P(\text{Features}|\text{Class}) \cdot P(\text{Class})}{P(\text{Features})}$$

Understanding the Dependency in Bayes' Theorem:

In Bayes' Theorem, we are working with the conditional probability:

$$P(\text{Red and Green}) = P(\text{Red}) \cdot P(\text{Green} \mid \text{Red})$$

The theorem itself doesn’t assume independence. It simply relates the prior probability of an event (how likely the event is before considering new evidence) and the likelihood (how likely the new evidence is, given that the event has occurred).

Now, let's break it down using your ball example and see how this ties into the Naive Bayes assumption of independence.

Where the Naive Assumption Comes In:

In Naive Bayes, we use Bayes' Theorem to calculate the probability of a class given some features. The naive part refers to the assumption that all features are independent. Let's say you have a classification problem where you're trying to predict whether an email is spam or not, based on several features (e.g., the presence of certain words like "free", "offer", etc.).

The probability of the class CCC (e.g., "spam" or "not spam") given the features X1,X2,…,Xn is:

$$P(C|X_1, X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n|C) \cdot P(C)}{P(X_1, X_2, \ldots, X_n)}$$

Now, instead of calculating the joint probability of all features together, we assume independence between the features:

$$P(X_1, X_2, \ldots, X_n|C) = P(X_1|C) \cdot P(X_2|C) \cdot \cdots \cdot P(X_n|C)$$

Conclusion:

You are correct that Bayes' Theorem itself deals with dependent events. However, in Naive Bayes, the assumption of independence simplifies the computation of the likelihood term. This is why, even though it's "naive" to assume that all features are independent, the method is often effective and computationally efficient.