Naïve Bayes

The Naïve Bayes algorithm is used for classification problems. It is highly used in text classification. In text classification tasks, data contains high dimensions (as each word represents one feature in the data). It is used in spam filtering, sentiment detection, rating classification etc.

This model predicts the probability of an instance belonging to a class with a given set of feature values. It is a probabilistic classifier. It is because it assumes that one feature in the model is independent of the existence of another feature. In other words, each feature contributes to the predictions with no relation between each other. In real world, this condition satisfies rarely. It uses Bayes theorem in the algorithm for training and prediction.

Bayes' Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$P(A|B) = P(B|A)\ P(A) P(B) P(A|B) = P(B) P(B|A)\ P(A)$

where A and B are events and $P(B) \neq 0$

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.
- P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).
- P(B) is Marginal Probability: Probability of **Evidence**.
- P(A|B) is a **posteriori** probability of B, i.e. probability of event after evidence is seen.
- P(B|A) is **Likelihood** probability i.e. the likelihood that a hypothesis will come true based on the evidence.

Advantages of Naive Bayes Classifier

- Easy to implement and computationally efficient.
- Effective in cases with many features.
- Performs well even with limited training data.
- It performs well in the presence of categorical features.
- For numerical features data is assumed to come from normal distributions

Disadvantages of Naive Bayes Classifier

- Assumes that features are independent, which may not always hold in real-world data.
- Can be influenced by irrelevant attributes.
- May assign zero probability to unseen events, leading to poor generalization.