Naive Bayes: Conditional Probability Basics

Conditional probability is used to calculate the probability of one event occurring given that another event has already occurred:

$P(A|B)$ reads as: “The probability of event A occurring given that B has already occurred.”

Conditional probabilities are calculated in the following way:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \quad ; \quad P(A|B)P(B) = P(A \text{ and } B) = P(B|A)P(A)$$

From there, Bayes’ Theorem can be derived:

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

Naive Bayes: The Central Assumption

Suppose we have k different variables $x_1, x_2, \ldots, x_k$ from which we want to predict a target class, Y. Bayes' theorem can then be used to find the conditional probability of a particular y given that $x_1, x_2, \ldots, x_k$ are true:

$$P(y|x_1, x_2, \ldots x_k) = \frac{P(x_1,x_2,\ldots,x_k|y)P(y)}{P(x_1,x_2,\ldots,x_k)}$$

The Naïve Bayes machine learning algorithms are based on Bayes' theorem with the assumption that all variables $x_1, x_2, \ldots, x_k$ are *independent*, meaning that no x value has any effect on the probability of another x value occurring. If we take all variables $x_1, x_2, \ldots, x_k$ as independent, then the following holds:

$$P(y|x_1, x_2, \ldots x_k) = \frac{P(x_1|y) * P(x_2|y) \cdots P(x_k|y) * P(y)}{P(x_1) * P(x_2) \cdots P(x_k)}$$

Naïve Bayes:

A Naïve Bayes algorithm finds the most likely class y for the given data points $x_1, x_2, \ldots x_k$ (the y value which maximizes the conditional probability equation, $P(y|x_1, x_2, \ldots x_k)$).

There are several different kinds of Naïve Bayes models, including:

1. Guassian Naïve Bayes- supports continuous values and assumes a normal distribution for each class.
2. Multinomial Naïve Bayes
3. Bernoulli Naïve Bayes

Advantages:

1) Simple to implement
2) Fast and effective in predicting the class of datasets
3) Doesn’t require a large test set to work well
4) Can handle both discrete and continuous data
5) Useful for text analysis and classification problems
6) Can be used for multiple class prediction problems

Disadvantages:

1) Relies upon the assumption that variables are independent
2) Not ideal for data sets with a large number of variables
3) Zero Frequency Problem

What data processing steps does the algorithm require

- **Convert categorical features to numeric (get dummies)**
- **Handle missing data. Drop or impute**
- **Identify outliers, remove or convert into numbers**
- **Separate dependent feature from independent features**
- **Gaussian- continuous data, need to be distributed normally**
- **Data smoothing- so that no probabilities are completely zero**

These are the most commonly adjusted hyperparameters with different Naive Bayes Algorithms

Gaussian Naive Bayes

- var_smoothing

Bernoulli Naive Bayes

- alpha
- fit_prior
- binarize

Multinomial and Categorical Naive Bayes

- alpha
- fit_prior

Complement Naive Bayes

- alpha
- fit_prior
- norm