Gaussian Naive Bayes

By: Savion Ponce, Juliann Groglio, Tavianne Kemp, Paul Polsinelli

How the Algorithm Works

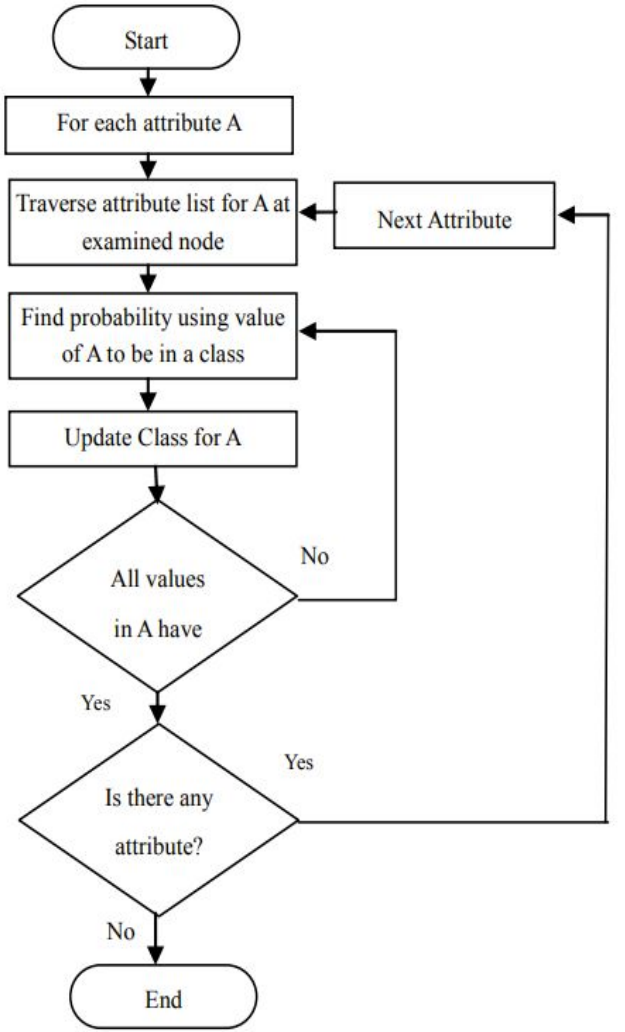
- The methods are a set of supervised learning algorithms.
- It uses the Bayes' theorem as a basis, Gaussian Naive Bayes theorem uses the equation below to calculate the conditional probability:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

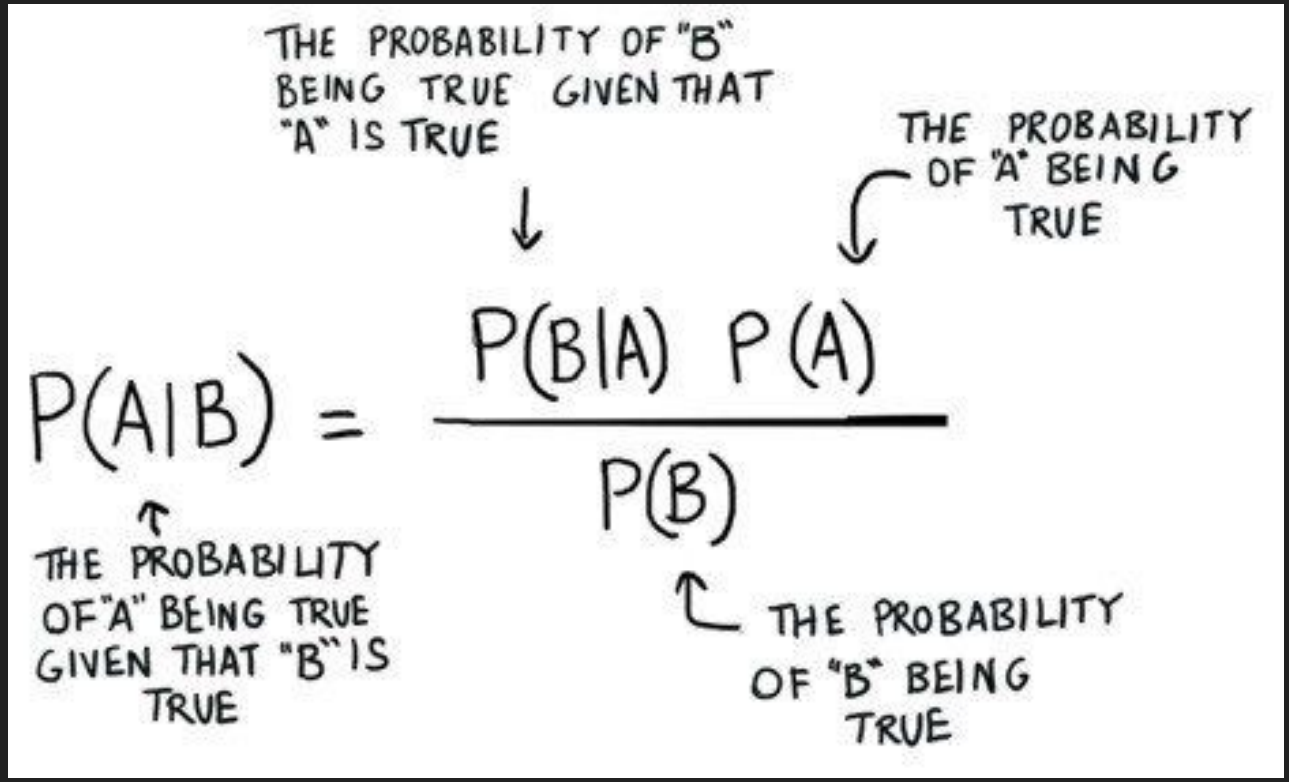
- A flowchart of the method is to the right:

Assumptions Made:

- **All Naive Bayes Algorithms:** All features are independent of one another. Meaning the presence of a feature will have no impact on the others and they are unrelated.
- **Gaussian**: Assumes that the continuous values corresponding to each feature are distributed according to Gaussian distribution, also called Normal distribution.



Flowchart of Naïve Bayes decision tree algorithm.


THE PROBABILITY OF "B" BEING TRUE GIVEN THAT "A" IS TRUE
THE PROBABILITY OF "A" BEING TRUE
P(A|B) = P(B|A) P(A) / P(B)
THE PROBABILITY OF "A" BEING TRUE GIVEN THAT "B" IS TRUE
THE PROBABILITY OF "B" BEING TRUE

Other types of Naive Bayes Models

- **Multinomial**: Frequencies of the occurrence of certain events represented by feature vectors are generated using multinomial distribution. This model is widely used for document classification.
- **Bernoulli**: In this model, the inputs are described by the features, which are independent binary variables or booleans. This is also widely used in document classification like Multinomial Naive Bayes.

Pros and Cons of Naive Bayes

Pros

- Easy to implement
- Fast
- Requires less training data
- Highly scalable
- Makes probabilistic predictions
- Robust
- Good with missing, continuous, and discrete data
- Easy to update

Cons

- Independent features not always truly independent
- If the category of a variable is not in the training set, it is assigned a 0 and no prediction can be made
- Estimations can be inaccurate in some cases. Not best for high accuracy cases

(Pedamkar, 2022)
(Shah, 2021)

Logistic Regression vs Naive Bayes

- Naive bayes is a generative model whereas Logistic Regression is a discriminative model. (Discriminative models draw boundaries in the data space, while generative models try to model how data is placed throughout the space. A generative model focuses on explaining how the data was generated, while a discriminative model focuses on predicting the labels of the data.)
- Naive bayes works well with small datasets, whereas Logistic Regression +regularization can achieve similar performance.
- Linear Regression performs better than Naive Bayes upon collinearity, as Naive Bayes expects all features to be independent.

(Shah, 2021)
(Goyal, 2021)

Processing Steps

Here are some useful processing steps to consider while implementing an NB algorithm:

- Import the libraries needed to run algorithm
- Import the dataset you would like to work with
- (Optional) Impute missing values or remove rows/columns if 75% of data is missing
- Splitting the dataset into training and test Set
- Feature Scaling

Just like in other algorithms learned, be sure to explore all other possible steps that could be particular to your data such as standardization and transformation. Also, NB can handle missing values, so the third bullet is optional.

Hyperparameters

- Hyperparameters for Gaussian NB is very limited. It is more of a generalized model
- Priors:
  - The probability of an event before new data is collected.
  - The best rational assessment of the probability of an outcome before an experiment is performed
  - If the priors are provided using an array, priors will not be adjusted based on the data
  - The basis for posterior probabilities
- Var_smoothing (Variance smoothing) :
  - Artificially adds a user-defined value to the distribution’s variance for calculation stability
  - If untouched, the default is 1e-9
  - This widens the curve and accounts for more samples that are further away from the distribution mean

Best Uses

- Text classification system
- Sentiment analysis
- Recommender system
- Real time prediction
- Multi-class prediction
- Spam filtering

Code Example

[1]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_iris
from sklearn.naive_bayes import GaussianNB
from sklearn import metrics

data = pd.read_csv('diabetes.csv')
```

[2]:
```
X = data.copy().drop(columns=['Outcome'])
y = data['Outcome'].copy()
```

[3]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=0)
```

[4]:
```
X_train.describe()
```

[5]:
```
X_test.describe()
```

[6]:
```
y_train.describe()
```

[7]:
```
y_test.describe()
```

Dataset Used: Pima Indians Diabetes Database | Kaggle
See our GitHub for the Jupyter NoteBook

Code Example Continued

[8]:
```
gnb = GaussianNB()
y_pred = gnb.fit(X_train, y_train).predict(X_test)
print("Number of mislabeled points out of a total %d points : %d"
% (X_test.shape[0], (y_test != y_pred).sum()))
```

```
Number of mislabeled points out of a total 384 points : 98
```

[9]:
```
model = GaussianNB()
model.fit(X_train, y_train)
print(model)
# make predictions
expected = y_test
predicted = model.predict(X_test)
# summarize the fit of the model
print(metrics.classification_report(expected, predicted))
print(metrics.confusion_matrix(expected, predicted))
print(model.score(X_train, y_train))
```

```
GaussianNB()
              precision    recall  f1-score   support

           0       0.78      0.86      0.82       253
           1       0.66      0.53      0.58       131

    accuracy                           0.74       384
   macro avg       0.72      0.69      0.70       384
weighted avg       0.74      0.74      0.74       384

[[217  36]
 [ 62  69]]
0.78125
```

Works Cited

Pedamkar, P. (2022, March 21). *Naive Bayes Algorithm*. EDUCBA. Retrieved July 21, 2022, from https://www.educba.com/naive-bayes-algorithm/

Shah, R. (2021). *Naïve Bayes Algorithm's Advantages and Disadvantages | Data Science and Machine Learning*. Kaggle. Retrieved July 21, 2022, from https://www.kaggle.com/getting-started/225022

Karim, M., & Rahman, R. M. (2013). Decision Tree And Naïve Bayes Algorithm For Classification And Generation Of Actionable Knowledge For Direct Marketing. *Journal of Software Engineering and Applications*, *06*(04), 196–206. https://doi.org/10.4236/jsea.2013.64025

Brownlee, Jason. "How to Prepare Data For Machine Learning." *Machine Learning Mastery*, 15 Aug. 2020, machinelearningmastery.com/how-to-prepare-data-for-machine-learning/#:%7E:text=The%20process%20for%20getting%20data%20ready%20for%20a,Step%201%3A%20Select%20Data%20Step%202%3A%20Preprocess%20Data.

Prabhakaran, Selva. "How Naive Bayes Algorithm Works? (With Example and Full Code) | ML+." *Machine Learning Plus*, 20 Apr. 2022, www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code.

Sharma, Mohit. "Data Preprocessing: 6 Necessary Steps for Data Scientists." *HackerNoon*, 27 Oct. 2020, hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing-502c993e1caa.

Works Cited Continued

Varghese, D. (2021, December 7). Comparative Study on Classic Machine learning Algorithms. Towards Data Science. Retrieved July 21, 2022, from https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222#:%7E:text=Naive%20bayes%20is%20a%20generative,all%20features%20to%20be%20independent.

Goyal, C. (2021, July 19). Deep Understanding of Discriminative and Generative Models in Machine Learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/#:%7E:text=Discriminative%20models%20draw%20boundaries%20in,the%20labels%20of%20the%20data.

sklearn.naive_bayes.GaussianNB. (n.d.). Scikit-Learn. Retrieved July 21, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

Can someone give a good math/stats explanation as to what the parameter var_smoothing does for GaussianNB in scikit learn? (2019, September 22). Stack Overflow. Retrieved July 21, 2022, from https://stackoverflow.com/questions/58046129/can-someone-give-a-good-math-stats-explanation-as-to-what-the-parameter-var-smoo