Naïve Bayes Algorithm

Instructor: Hamed Haheminia

Lecture 17

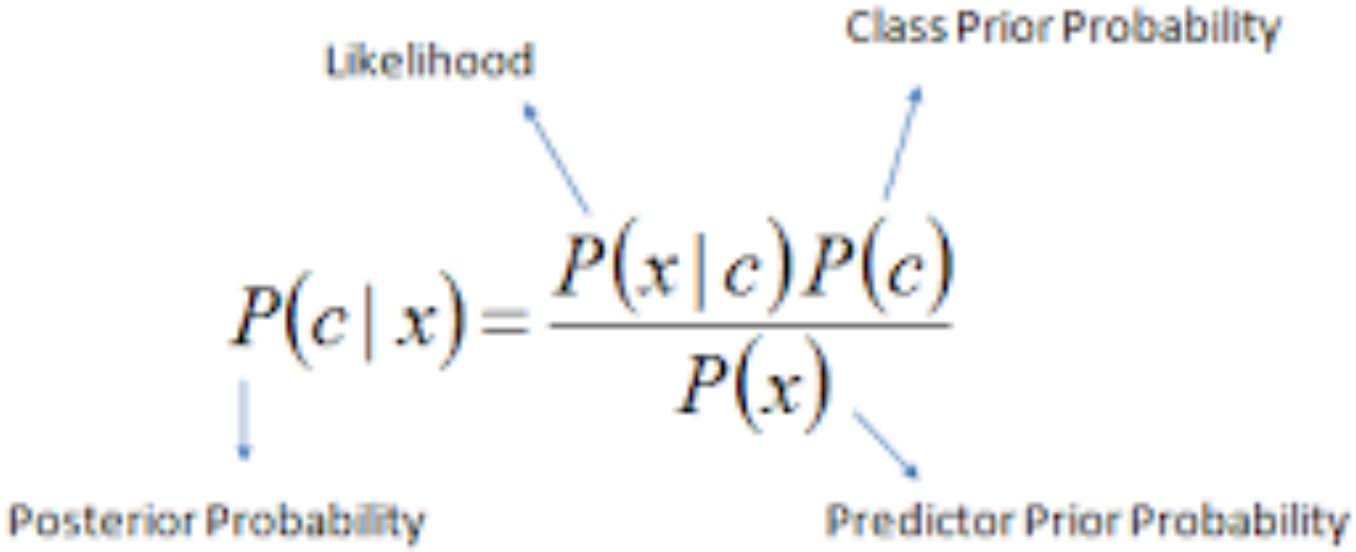
Agenda

- What are Naïve Bayes Algorithms?
- How NB works?
- Advantages and disadvantages of using NB
- Gaussian, Multinomial, and Bernoulli NB models.

Naïve Bayes Algorithms

- Naïve Bayes algorithms are among the most famous supervised learning algorithms.
- They are used in classification.
- NB is inherently multiclass – i.e. you can easily apply it to cases where you have more than one type of output.
- There are extremely easy to build and very useful for very large data sets.

Bayes Theorem



Naïve Bayes

- Naïve Bayes Assumption:
  - **Features are independent** given class (This is a strong assumption):
  - $P(X_1,X_2|c_j) = P(X_1|X_2, c_j)P(X_2|c_j) = P(X_1|c_j)*P(X_2|c_j)$

$$P(x_1, x_2, \ldots, x_n \mid c_j) = \prod_i P(x_i \mid c_j)$$

$$c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_i P(x_i \mid c_j)$$

Example. 'Play Tennis' data

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Question: For the day <sunny, cool, high, strong>, what's the play prediction?

Solution

$$P(PlayTennis = yes) = 9/14 = 0.64$$

$$P(PlayTennis = no) = 5/14 = 0.36$$

$$P(Wind = strong \mid PlayTennis = yes) = 3/9 = 0.33$$

$$P(Wind = strong \mid PlayTennis = no) = 3/5 = 0.60$$

etc.

$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$$

$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = \mathbf{0.0206}$$

$$\Rightarrow answer : PlayTennis(x) = no$$

Underflow Prevention

- Multiplying lots of probabilities, which are between 0 and 1 by definition, can result in floating-point underflow.
- Since *log(xy) = log(x) + log(y),* it is better to perform all computations by summing logs of probabilities rather than multiplying probabilities.
- Class with highest final un-normalized log probability score is still the most probable.

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$

In-class practice

- Use 1 nearest neighbor algorithm and locate the closest classmate to you. Once you locate him/her, discuss how one could use Naïve Bayes classifier to classify Spam and Ham e-mail. Also, discuss how we can measure the error of this algorithm and what can go wrong.

A few issues with Bayes Algorithm

- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability to it and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- This is parameter “alpha” in BernoulliNB and MultinomialNB in python. Default is “alpha = 1”
- Naive Bayes a bad estimator – don’t take probability estimations seriously!

Different Types of NB models in Python

- What does dictate the type of NB model you can use? The answer is your feature types. Depending on features, you can use 3 different NB models: Gaussian, Bernoulli, and Multinomial
  - If feature space is quantitative then you shall use GaussianNB
  - If feature space is Binary, then you better use BernoulliNB
  - If feature space is discrete counts, then you can use MultinomialNB.
  - Can you come up with few examples for each class of NB?

Gaussian NB

- Your assumption for Gaussian NB is, your feature variables are independent and Normally distributed.
- You should make sure, your input features either look normal at their raw format, or look normal after transformation. For instance you can use log transform to make most of positively skewed distributions, symmetric.
- For Gaussian NB, we first need to estimate the mean and Standard deviation of each feature.
- Once you successfully calibrate your model, you can use probability density function of Normal Distribution to calculate likelihood functions.

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

Advantages of using NB algorithm

- Very easy to compute
- Great for multi-class cases
- One of the fastest algorithms
- There is no parameter needed to be tuned. (an exception might be alpha – which by definition is not a tuning parameter).
- The algorithms can be used for **real time** prediction.
- Used often in Text Classification/Spam Filtering/Sentiment Analysis.

Issues with NB

- It works under the strong assumption that your feature inputs are independent.
- If you have highly dependent variables, you must drop one.
- GaussianNB works under the assumption that your inputs are normally distributed. If that is not the case, you either cannot use it or need to transform your variables.
- You cannot take probability predictions seriously!

Summary

- Definition of NB algorithm
- Learned how NB algorithm works
- Gaussian NB, Multinomial NB and Bernoulli NB
- Limitations and advantages of NB