Bayesian Statistics in machine learning

A Scientific Research Paper

Submitted to Professor: Frank Zimmer

Quang Minh, Dinh( 24211)

**Abstract**

Bayesian statistics is a system that describes epistemological uncertainty using the mathematical probability language. In the Bayesian paradigm degrees of belief in nature states are specified; they are not negative and the total belief in all states of nature is fixed as one. Bayesian statistical methods start with existing previous beliefs and update those using data to give back beliefs that can be used as the basis for inferential decisions. Bayesian statistics provides a principled way for calculating a conditional probability. It is a deceptively simple calculation, although it can be used to easily calculate the conditional probability of events where intuition often fails. Although it is a powerful tool in the field of probability, Bayesian statistics is also widely used in the field of machine learning. In this scientific paper, we will discover Bayesian statistics for calculating conditional probabilities and how it is used in machine learning.

**Introduction**

Bayes’ theorem

In probability theory and statistics, Bayes’ theorem, named after Reverend Thomas Bayes, describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if the risk of developing health problems is known to increase with age, Bayes’ theorem allows the risk to an individual of a known age to be assessed more accurately than simply assuming that the individual is typical of the population as a whole.When applied, the probabilities involved in the theorem may have different probability interpretations. Bayesian inference is fundamental to Bayesian statistics.

Machine learning

Machine learning (ML) is the study of innovative computer algorithms that automate through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build product code based on sample data, known as "training data", to make predictions or decisions without explicit programming. Machine learning algorithms are used in many different applications, such as Email filtering and computer vision, where it is difficult or impossible to develop conventional algorithms to perform tasks requirement.

**State of Bayes’ theorem**

Bayes’ theorem is stated mathematically as the following equation (simple form):

Where and are events and  **(see figure 1).**

1. {\displaystyle P(A\mid B)} is a conditional probability: the likelihood of event occurring given that is true.
2. is a conditional probability: the likelihood of event occurring given that is true.
3. and are the probabilities of observing andrespectively; they are known as the marginal probability.
4. and must be different event.

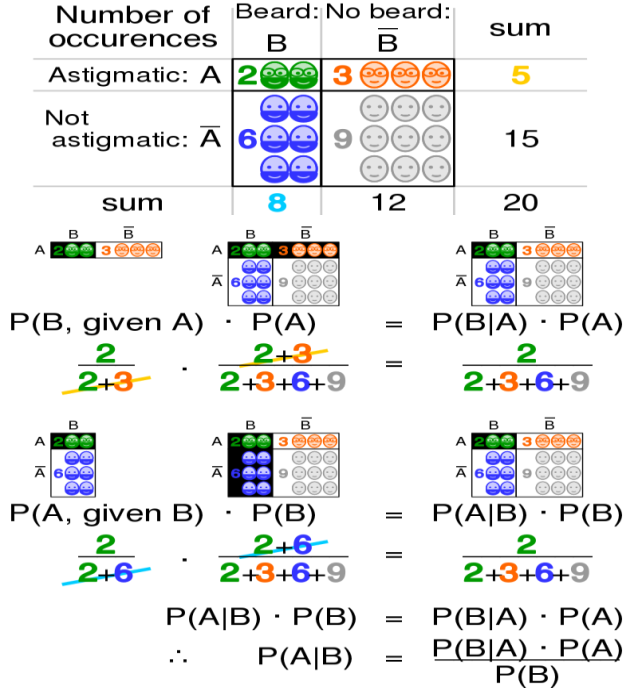


Figure 1: A geometric visualization of Bayes’ theorem.

**Bayesian Approach.**

In Bayesian machine learning we use the Bayes rule deduce model parameter (*θ)* from data (D):

is a normalizing constants and stay the same for each model.

is a prior, or what is our belief about the model parameters.

is likelihood of data given model parameters.

is a posterior and a probability distribution over model parameters obtained from prior beliefs and data.

The Bayesian Approach to Machine Learning

1. Formulating our knowledge about the problem probabilistically:

* Define a model that expresses qualitative aspects of our knowledge.
* Specify a prior probability distribution for these unknown parameters that express out beliefs about which values are more or less likely, before seeing the data.

1. Gathering the data.
2. Computing the posterior probability distribution for the parameters, given the observed data.
3. Using this posterior distribution to:

* Reach scientific conclusion.
* Make predictions by averaging over the posterior distribution.
* Make decisions so as to minimize posterior expected loss.

**Example of Bayes’ Theorem**

**Example: Bayes –Dice**

“You have a fair dice A and an unfair dice B and a long sequence S of observations. Using Bayes’ theorem, how can you remove whether the sequence was made by dice A or by dice B?”

The following dice are given:

Fair dice A:

Unfair dice B:

---------------------------------------------------

**First step (Function):** Complete the fair dice A and the unfair dice B function

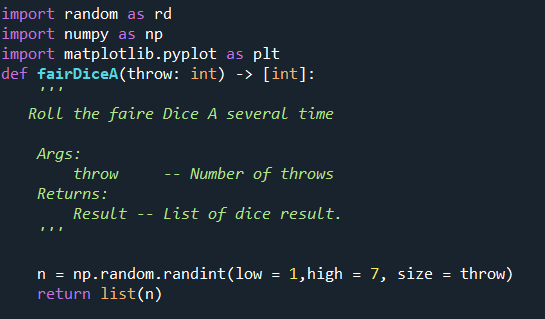


Figure 2: Fair dice A function.

Explanation: Fair dice A, all numbers (1-6) have equal probability ( so we can implement by random function with no difference between numbers.

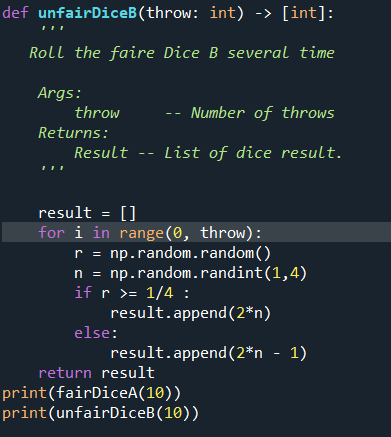


Figure 3: Unfair Dice B function.

Explanation: Unfair dice B, with the difference between odd and even numbers, an integer number ( was declared to simplify the probability them. If variable random , results will be even numbers(2,4,6 ↔ 2\*n), otherwise odd number will be represented( 1,5,7↔2\*n-1).

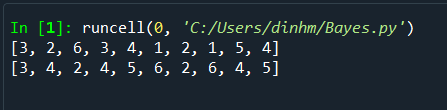
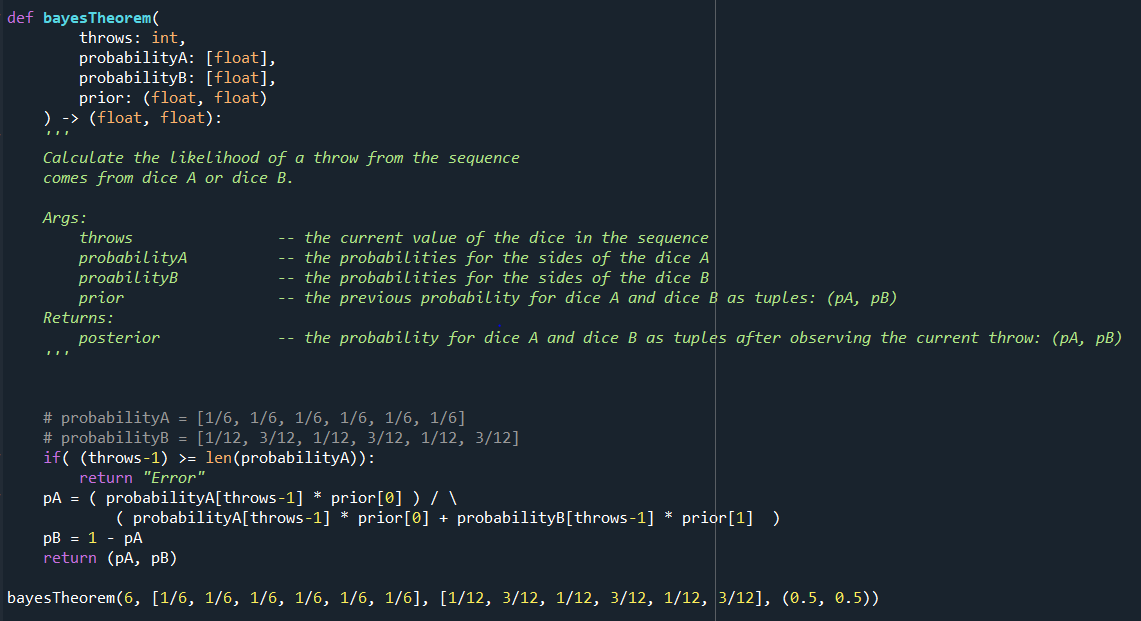


Figure 4: Result of the first step.

**Second Step (Bayes Theorem):** Implement the probability distributions for throwing the dice of the two dice and the previous probability for dice A and B, the new probabilities for dice A and B are calculated.

Figure 5: Bayes’ theorem method.

Explanation: In the current sequence, with Bayes’ theorem denominator is the probability of value 6(dice A throws value 6 and dice b throws value 6) (line 75 figures 5) and nominator is the probability of dice A throws value 6.

Probability of dice A throwing value 6 is the probability of value 6 of dice A multiplies the previous prior A.

**Third step (Visualization graph):**The following sequence of dice rolls of a dice (fair or unfair) is given: Sequence= [1, 1,. . . , 2, 2,. . . , 3, 3,. . . , 4, 4,. . . , 5, 5,. . . , 6, 6] The sequence consists of the following: 10 times a 1, 10 times a 2, 10 times a 3, 10 times a 4, 10 times a 5 and 10 times a 6. Iterate this sequence with the function from subtask for the dice A and B and find the probability that A or B is present. Plot the result for dice A and B after each roll.

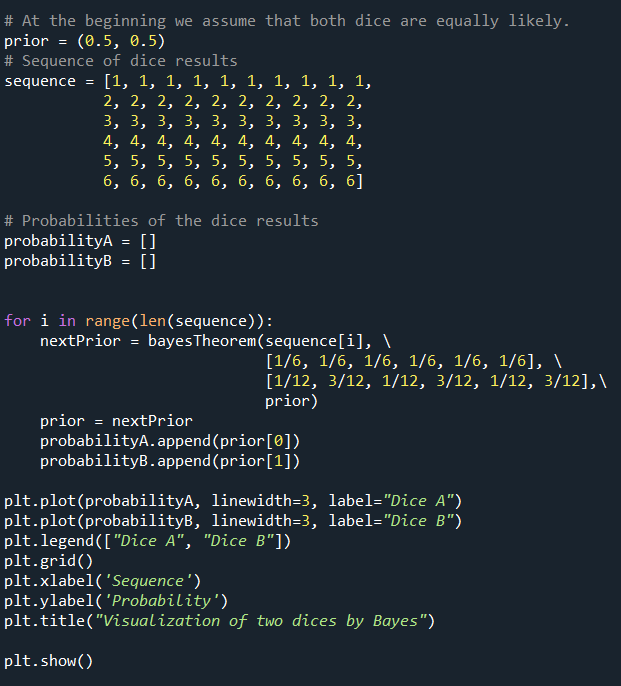


Figure 6: Bayes’s theorem with given sequence

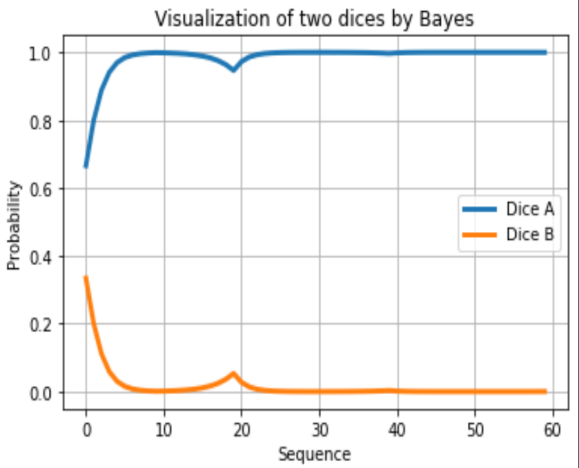


Figure 7: Visualization of two dice by Bayes.

**Discussion and conclusion**

Considering the first ten times, the probability that dice B decreases, the probability of dice A increases because we get ten numbers 1 where dice B has the probability of odd numbers smaller than dice A. Similarly, after the next 10 attempts, we got ten times value of 2, so the probability of dice B increases. The reason is the probability of even value (dice B) is bigger and that causes a small rise around twentieth attempt (see figure 7). To roll a dice ten times of value 3, this probability is the same as the first ten times. But starting from the thirtieth, however, the graph does not change due to Bayes' theorem.

* Bayes’ theorem is a causal system that depends on previous prior.

**Application of Bayesian Statistics**

Explicitly, Bayesian statistical methods tend to be used in three main situations. The first is a situation where one has no alternative but to include quantitative previous judgments, due to a lack of data on some aspects of the model or due to a lack of some evidence to be recognized through assumptions about bias. These situations can arise when a political decision has to be made on the basis of a combination of imperfect evidence from different sources, such as the support of Bayesian methods by the Food and Drug Administration responsible for medical devices. Common areas of application include meta-analysis, disease mapping, multi-center studies, etc. With poorly informative prior distribution, conclusions can often be numerically similar to classical techniques, although interpretations may differ. The third area concerns the construction of a huge common probability model that combines perhaps thousands of observations and parameters, and the only feasible way to derive unknowns is to use the Bayesian approach: examples include image processing, spam filtering, signal analysis, and data gene expression. There is also extensive use of Bayesian ideas of parameter uncertainty but without the explicit use of Bayes’ theorem. If a deterministic prediction model has been constructed, but some parameter inputs are uncertain, a common previous distribution can be placed on these parameters and the resulting uncertainty propagated through the model, often using Monte Carlo methods, to create a predictive probability distribution. This technique is widely used in risk analysis.

References

* [Mitchell, Tom](https://en.wikipedia.org/wiki/Tom_M._Mitchell) (1997). [Machine Learning](http://www.cs.cmu.edu/~tom/mlbook.html). New York: McGraw Hill. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)) [0-07-042807-7](https://en.wikipedia.org/wiki/Special:BookSources/0-07-042807-7). [OCLC](https://en.wikipedia.org/wiki/OCLC_(identifier)) [36417892](https://www.worldcat.org/oclc/36417892).
* A Gentle Introduction to Bayes Theorem for Machine Learning by [Jason Brownlee](https://machinelearningmastery.com/author/jasonb/).
* Joyce, James (2003), ["Bayes' Theorem"](https://plato.stanford.edu/archives/spr2019/entries/bayes-theorem/), in Zalta, Edward N. (ed.), The Stanford Encyclopedia of Philosophy (Spring 2019 ed.), Metaphysics Research Lab, Stanford University, retrieved 2020-01-17
* Stuart, A.; Ord, K. (1994), Kendall's Advanced Theory of Statistics: Volume I—Distribution Theory, [Edward Arnold](https://en.wikipedia.org/wiki/Edward_Arnold_(publisher)), §8.7
* [Mitchell, Tom](https://en.wikipedia.org/wiki/Tom_M._Mitchell) (1997). [Machine Learning](http://www.cs.cmu.edu/~tom/mlbook.html). New York: McGraw Hill. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)) [0-07-042807-7](https://en.wikipedia.org/wiki/Special:BookSources/0-07-042807-7). [OCLC](https://en.wikipedia.org/wiki/OCLC_(identifier)) [36417892](https://www.worldcat.org/oclc/36417892)..
* [Bishop, C. M.](https://en.wikipedia.org/wiki/Christopher_M._Bishop) (2006), Pattern Recognition and Machine Learning, Springer, [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)) [978-0-387-31073-2](https://en.wikipedia.org/wiki/Special:BookSources/978-0-387-31073-2).
* Machine learning and pattern recognition "can be viewed as two facets of the same field."[[3]](https://en.wikipedia.org/wiki/Machine_learning#cite_note-bishop2006-3):vii
* [Friedman, Jerome H.](https://en.wikipedia.org/wiki/Jerome_H._Friedman) (1998). "Data Mining and Statistics: What's the connection?". Computing Science and Statistics. 29 (1): 3–9.
* Ethem Alpaydin (2020). Introduction to Machine Learning (Fourth ed.). [MIT](https://en.wikipedia.org/wiki/MIT). pp. xix, 1–3, 13–18. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)) [978-0262043793](https://en.wikipedia.org/wiki/Special:BookSources/978-0262043793).
* Gelman, Carlin, Stern and Rubin (2003) Bayesian Data Analysis (Second Edition) Chapman and Hall/CRC.
* Bernardo and Smith (1994) Bayesian Theory, Wiley.

## Where priors come from? (**Rob Zinkov** 2015-06-09).