**BAYESIAN LEARNING**

**by**

**Reethika Reddy Challa**

**Meghana Akshitha Ponnapalli**

**Vishwas Sabbani**

**DS 802: Probability and Simulation**

**University of New Hampshire**

**Abstract:**

This study provides an overview of the fundamentals of Bayesian learning methods in the context of machine learning, with a focus on the significance of marginalization for dealing with the uncertainty. It provides more details on what Bayesian learning is, its methodologies, such as Maximum Likelihood and Maximum A Posterior Hypothesis, and its applications. It is a method for choosing between conflicting hypotheses. The goal is to determine the best hypothesis based on the data sample with the best hypothesis being the most plausible one once all the real data has been gathered. These approaches remain among the most useful ones for specific kinds of learning procedures. Example: Calculate the probability of a hypothesis using Baye's classifier. They offer an insightful viewpoint for comprehending numerous learning algorithms that do not explicitly manipulate probabilities.

**The History of Bayes’ Learning:**

Bayes theorem was named after Thomas Bayes, who was an English statistician, philosopher, and Presbyterian minister in England in the early 18th century. The fact is that he never really came up with any formula related to Bayes theorem, but he was elected as a Fellow of the Royal Society based on his work and publications on Newton's calculus. His friend Richard Price published his work on inverse probability after he died. Inverse probability is the philosophical basis for Bayesian statistics. He was elected to the Royal Society after he published this work. However, the actual formula was published by Pierre-Simon Laplace for his work on inductive reasoning, which is based on probability.

While researching the history of Bayes, we came across an intriguing story about how Bayes used to perform an experiment with a ball to calculate the probability of the ball landing in a specific position.Bayes asked his assistant to drop a ball while he sat with his back against a desk. Whenever a ball is dropped, it could go anywhere based on the level of the desk. Bayes wants to hypothesize the probability of the ball going in a particular direction. He considers the ball in a particular position as its original position. He asked to drop the ball several times to the right or left, further away or further toward him. He asked me to do this several times. As this experiment was done several times, he got more information on the position of the ball. The more information he gets, the more accurate his predictions can be on the first ball position. He was just taking his initial belief and adding more information to get an accurate belief.

**What is Bayesian learning?**

Bayesian learning is a learning technique that determines model parameters by maximizing the posterior probability of the parameters given the training data. Some parameter values are supposed to be more in line with the observed data than others. According to Bayes' rule, increasing the posterior probability equates to increasing the "model evidence," which is the conditional likelihood of the training data given the model parameters. It is frequently possible to estimate the evidence using a closed formula or an updating rule. Instead of saving patterns for parameter cross-validation, Bayesian approaches enable the use of all data for training. It determines how concepts are modeled using probability.

Combining the information we have already seen with our preexisting opinions is a potent machine learning technique. It blends the likelihood of the facts with what you already know. In general, it is a technique for choosing the best theory in terms of how well it can explain the training data that has been observed.

Examples: detecting a disease, diagnostic results, texting suggestions on the phone, speech recognition, etc.

**Example:**

Consider a scenario in which your friend hands you a fresh coin and asks you to judge its fairness (or the likelihood of seeing heads) before even flipping it once. You are even aware that your acquaintance did not intentionally prejudice the coin. You know that coins are generally fair, so you anticipate that the likelihood of seeing heads will be 0.5. You can only use your prior experiences or observations with coins to claim the fairness of the coin in the absence of any such observations.

Assume you are given permission to flip the coin ten times in order to determine its fairness. One of the following categories will apply to your experiment-related observations:

Case 1: Observing 5 heads and 5 tails in Case 1

Case 2: seeing 10-h tails with 10-h heads.

You will be more certain that the coin is fair and will determine that there is a 0.5 chance of seeing heads if instance 1 is observed. You have two choices if case 2 is observed:

Since you now have new information, disregard your previous assumptions and determine that the likelihood of seeing heads is h/10 entirely based on recent observations.

Adapt your belief to the value of h that you just learned.

The first option recommends that we make decisions without taking our beliefs into account, or the frequentist method. Ten coins is insufficient to judge a coin's fairness, so the second way appears to be more practical. Therefore, by combining our most recent observations with the assumptions we have developed from our prior experiences, we can make better decisions. Bayesian thinking is a way of thinking that makes use of our most recent observations as well as our beliefs or propensity for critical thought.

Furthermore, suppose your friend agrees to let you carry out an additional 10 coin tosses. Then, we can further revise our views using these fresh observations. We can gradually change our ideas as new information comes in, making our judgments more certain. You update your understanding gradually with fresh data in a process known as incremental learning.

When frequentist statistics cannot be used in these situations because of the drawbacks we have already described, Bayesian learning is used. We can use Bayesian learning to solve all of these problems when attempting to estimate unknown parameters for machine learning models, and even with additional capabilities (such as incremental updates of the posterior). With some supporting data or observations, Bayesian learning uses the Bayes theorem to calculate the conditional probability of a hypothesis.

**Features of Bayesian learning methods:**

These methods enable the estimated likelihood that a hypothesis is true to be steadily decreased or increased for each observed training example. Compared to algorithms, which discard a hypothesis if it is discovered to be inconsistent with any particular example, this technique for learning is more flexible. It just modifies the theory rather than rejects it. The final probability of a hypothesis can be calculated using observed data and prior knowledge. For each candidate hypothesis in Bayesian learning, a prior probability is asserted, along with a probability distribution across the observed data, to represent prior knowledge. By aggregating the predictions of many hypotheses and weighting them according to their probabilities, new occurrences can be classified using Bayesian approaches to account for hypotheses that make probabilistic predictions. Even in situations when Bayesian procedures are computationally infeasible, they can offer a benchmark for the best possible decision-making that can be used to compare alternative realistic approaches. Machine Learning Two Bayesian methods Difficulties basic understanding of many probabilities When these probabilities are unknown beforehand, they are frequently calculated using prior information, data that was previously available, and presumptions regarding the shape of the underlying distributions. In the general scenario, determining the Bayes optimum hypothesis is computationally expensive (linear in the number of candidate hypotheses). This processing cost can be greatly decreased in some unique circumstances.

**Baye’s theorem:**

A mathematical formula for determining the likelihood of an outcome, based on a previous outcome having occurred in similar circumstances(conditional probability). Bayes' Theorem allows you to update the predicted probabilities of an event by incorporating new information. Bayes' theorem relies on incorporating [prior probability](https://www.investopedia.com/terms/p/prior_probability.asp) distributions in order to generate [posterior probabilities](https://www.investopedia.com/terms/p/posterior-probability.asp).

**Prior probability:**

The chance of an event occurring before fresh data are gathered is known as the prior probability in Bayesian statistical reasoning. In other words, it represents the most logical estimate of the likelihood of a specific event based on the available information prior to the execution of an experiment.

**Posterior probability:**

The probability of an event occurring with the additional information taken into account is known as the posterior probability. By utilizing Bayes' theorem to update the prior probability, posterior probability is determined. The posterior probability, as used in statistics, is the likelihood that event A will take place after event B.

**Example:**

Consider a random individual is selected from a population which consists of two ethnic groups, Group A and Group B.

Given below is some information we know:

* 30% of individuals belonging to group ABC have incomes below the poverty line;
* the corresponding proportion for the population as a whole is 20%;
* 40% of the population is made of individuals belonging to group ABC

From the given, we can say that

The prior probability is P(A), the probability of belonging to group A.

The posterior probability P(A|poor) can be computed using Baye’s theorem.

**Formula:**

P(h|D) = P(D|h)

P(h) P(D)

* P(h) stands for the prior probability of the hypothesis h.
* P(h), which stands for the initial likelihood that hypothesis h is true, prior to training data observation.
* P(h) may reflect any prior knowledge we have regarding the likelihood that h is accurate. Without this prior information, each contender theory might just receive the same prior probability.
* P(D) stands for the prior probability of the training data.
* The likelihood of D given that it is unknown which hypothesis is correct.
* P(h|D) stands for the probability of h given D.
* P(h|D) is referred to as the posterior probability of h since it expresses how confident we are that h is true after seeing the training data D.
* The posterior probability P(h|D) reflects the influence of the training data D, in contrast to the prior probability P(h), which is independent of D.
* P(D|h) = probability of D given h
* The probability of observing data D given some world in which hypothesis h holds.
* Generally, we write P(xly) to denote the probability of event x given event y.

**Example:**

Suppose that a test for using a particular drug is **97% sensitive** and **95% specific**. That is, the test will produce **97% true positive** results for drug users and **95% true negative results** for non-drug users. These are the pieces of data that any screening test will have from their history of tests. We also know that 0.5% of the general population are users of the drug. What is the probability that a randomly selected individual with a positive test is a drug user?

P(user) = 0.005

P(non-user)= 1-0.005 = 0.995

P(+ve | user) = 0.97

P(-ve | user) = 1-0.97= 0.03

P(+ve | non-user)= 0.95

P(-ve | non-user) = 1-0.95= 0.05

Using Baye’s theorem

P(user | +ve) = P(+ve | user) P(user)

P(+ve)

P(+ve) = P(+ve | user) P(user) + P(+ve | non-user) P(non-user)

P(user| +ve)= 0.089

This shows that the total number of positively tested are almost 89% among the users.

**Maximum Likelihood**

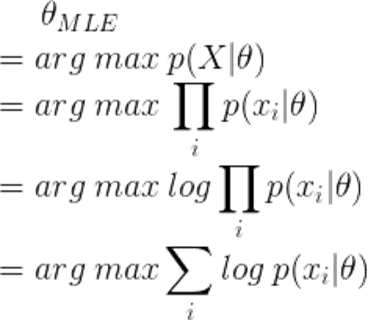
Maximum Likelihood is a specific instance of MAP from which the probabilities of all candidate hypothesis are equal.

**hML = argmax (h € H) P(D|h)**

This theory increases the P(D|h).

The parameters of a distribution are computed using the Maximum Likelihood Estimation (MLE) method. The parameters of machine learning models like Naive Bayes and Logistic Regression are routinely calculated using MLE as well. Because it is so common and well-liked, some people use MLE even if they are unfamiliar with it. For instance, one can rapidly compute the sample mean and variance and use these as the distribution's parameters for fitting a normal distribution to a dataset.

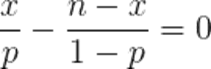
The goal of MLE is to infer Θ in the likelihood function p(X|Θ).



In the equation, xi denotes a single trail (0 or 1) and x denotes the overall number of heads. Take a log for probability after that:



If we take the log likelihood function's derivative in relation to p, we get:



Finally, the estimate of *p* is:



As a result, in this instance, 0.7 of the time a conventional coin will land on its head. This coin is obviously unfair.

The parameter p's maximum likelihood estimate (MLE) is the value of p that maximizes the likelihood P(data | p) given the data. In other words, the MLE is the p value for which the data is most likely to exist.

The empirical probability of success in a sequence of Bernoulli trials will converge to the theoretical probability in accordance with the law of large numbers. But if you flip this coin ten times, you'll get seven heads and three tails. What about the conclusion?





We cannot exclude the probability that p(Head) = 0.5 even if p(Head = 7| p=0.7) is bigger than p(Head = 7| p=0.5). That is the issue with MLE (Frequentist inference). It never provides or utilizes a hypothesis' probability.

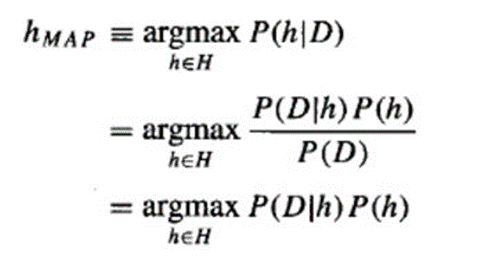
**Maximum A Posterior Hypothesis**

Using a model and a prior probability or belief about the model, MAP entails computing a conditional probability of observing the data.

In light of the observed data D, a learner considers a number of possible hypotheses H and is interested in determining which one is the most likely.

The term "MAP" refers to any such maximum probable hypothesis.

By calculating the posterior probability of each candidate hypothesis, we may identify the MAP hypothesis using the Bayes theorem.



MAP typically surfaces in a Bayesian context. It works on a posterior distribution, not only the likelihood, as the name implies.

An optimization problem for determining the posterior probability's central tendency is solved by maximizing this value over a range of theta (e.g. the model of the distribution). As a result, this method is also known as "maximum a posteriori estimation," also known as "maximum posterior estimate" or "MAP estimation."

Fundamentally, the Bayesian method of statistics places a strong emphasis on using all available data to deduce conclusions in the face of uncertainty. These facts may come from recently discovered information, research that has already been done, both, as is usually the case, or neither.

What is a prior distribution, exactly? Our parameter of interest,, has a prior distribution that we denote as P(θ); this function indicates which values of are more or less likely, according to our interpretation of previously available relevant data. The likelihood function, proportional to P (D | θ ), which is then multiplied by the prior distribution (and rescaled) to create the posterior distribution, P ( θ | D), with which we can carry out our desired inference, represents the knowledge obtained from our new data. The likelihood function, then, acts to convert our prior distribution into a posterior distribution, as a result.

Mathematically, we use Bayes’ rule to obtain the posterior distribution of our parameter of interest θ,

**P(θ | D) = K × P(θ) × P(D | θ)**,

where in this context K = 1/P(D) is merely a rescaling constant. We often write this more simply as

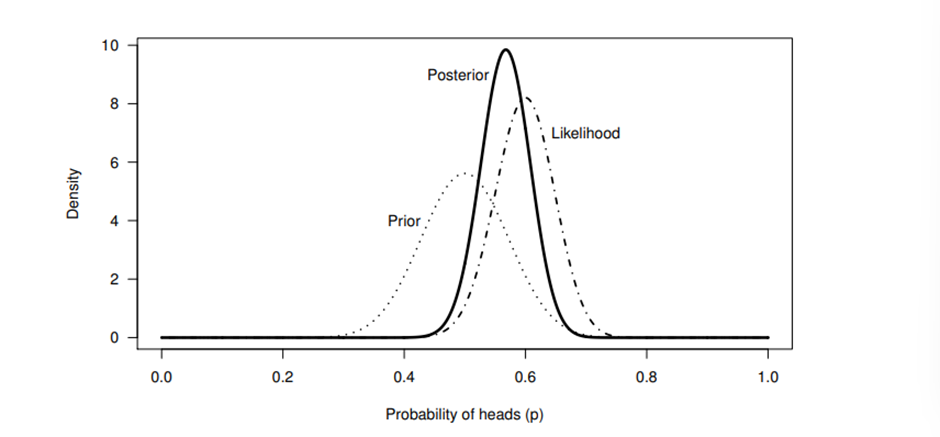
**P(θ | D) ∝ P(θ) × P(D | θ),**

where ∝ means “is proportional to”. In words, we say the posterior distribution is proportional to the prior distribution multiplied by the likelihood function.

Think back to the coin example from earlier, where 60 out of 100 flips resulted in heads. It is likely that p is in the range of.30 to.70 if we had some reason to believe that the coin's bias was within.2 of being fair in either way before to the experiment. We could opt to use the Beta(25,25) distribution to visualize this data. Since the dot-and-dashed line representing the probability function for the 60 flips is identical to the one in the center panel, we can be certain that the solid line representing the Beta(85,65) posterior distribution will represent the experiment's results.

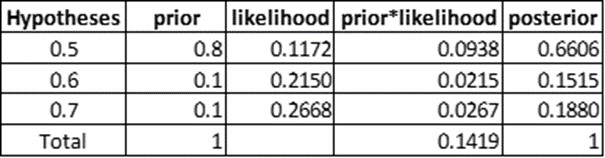
The solution to our Bayesian estimation problem is represented by the whole posterior distribution, although we frequently offer summary measures to make the results' communication easier. For instance, as a best guess for the value of p, we may choose the maximum of the posterior distribution, or maximum a posteriori (MAP) estimate, which in this case is pmap =.568.

Notice that this is slightly different from the estimate we would obtain if we computed the maximum likelihood estimate, pmle = .60. This discrepancy is due



**Example:**

Prior probabilities are first assigned values in a Bayesian study. Three possibilities are presented here: p(head) equals 0.5, 0.6, or 0.7. Prior probability that match are 0.8, 0.1, and 0.1.



Because the likelihood is now weighed by the prior, even while the likelihood reaches its peak in this case at p(head)=0.7, the posterior peaks at p(head)=0.5. Inferred from MAP, p(Head) = 0.5. We might arrive at a different conclusion if the prior probability in column 2 were altered, though. As a result, a subjective prior is, well, subjective, which is one of the fundamental criticisms of MAP (Bayesian inference).

**Example:**

100 out of 10,000 women aged forty who participate in a routine screening have breast cancer

80 of every 100 women with breast cancer will get positive tests

950 out of 9,900 women without breast cancer will also get positive tests

**PROBLEM**

What percentage of women with positive tests who are in this age group and have a routine screening will actually have breast cancer?

100 breast cancer patients were present prior to the screening

9,900 non-breast cancerous females

Following the screening:

A = 80 women who tested positive for breast cancer.

B = 20 breast cancer patients with negative test results

C = 950 women with positive test results but no breast cancer

D = 8,950 women who did not have breast cancer and tested negative

Within the category of ALL patients who received favorable results, the percentage of cancer patients:

A/(A+C) = 80/(80+950) = 80/1030 = 0.078 = 7.8%

Prior Probabilities:

100/10,000 = 1/100 = 1% = p(C)

9,900/10,000 = 99/100 = 99% = p(~C)

Conditional Probabilities:

A = 80/10,000 = (80/100)\*(1/100) = p(T|C)\*p(C) = 0.008

B = 20/10,000 = (20/100)\*(1/100) = p(~T|C)\*p(C) = 0.002

C = 950/10,000 = (9.6/100)\*(99/100) = p(T|~C)\*p(~C) = 0.095

D = 8,950/10,000 = (90.4/100)\*(99/100) = p(~T|~C) \*p(~C) = 0.895

Rate of cancer patients with positive results, within the group of ALL patients with positive results:

P(C|T) = =

A/(A+C) = 0.008/(0.008+0.095) = 0.008/0.103 = 0.078 = 7.8%

**Applications:**

**Credit card fraud detection:** By analyzing data and estimating probabilities using Bayes' theorem, We can identify trends and clues to detect credit card fraud. Credit card fraud detection can lead to false positives due to insufficient data. A Bayesian Neural Network algorithm is applied to customer profile records, including each customer's past financial transactions, when unexpected behavior is reported to enterprise risk management. These analyzes verify the presence of indicators of fraud.

**Spam filtering:** Bayesian inference can identify spam messages by applying Bayes' theorem to create a model that can determine if an email is likely to be spam. Each word in the message is considered by a Bayesian model trained using Bayesian techniques and given a variable weight based on how often it occurs in spam and non-spam communication. To determine if an email is spam, Bayesian Neural Networks analyze variables such as word count, word length, and the presence or absence of certain characters.

**Medical diagnosis:** Bayes' theorem is used in medical diagnostics to predict a patient's likelihood of contracting a particular disease using data from previous cases. Bayesian inference allows better predictions than traditional statistical techniques because it can consider all variables that can affect the outcome and provides probabilities, not just binary outcomes. To assess disease likelihood, posterior probabilities are calculated from Bayes' theorem and linked to clinical information about disease and symptoms. Uses Bayesian inference to make diagnoses by examining historical patient data and identifying patterns that help determine whether a patient has Alzheimer's disease. Bayes' theorem is particularly useful when making predictions for rare diseases when you need data for .

**Support robot decisions:** Bayesian inference is used in robotics to support robot decision making. Using real-time sensor data from the robot's environment, Bayes' theorem can be used to predict the robot's next move or action based on past performance. The robot uses Bayes' theorem to extract relevant data from the environment, such as direction of movement, speed, and obstacles. Bayesian reinforcement learning can be used to train robots. Bayesian Reinforcement Learning (BRL) uses previous experience, knowledge, and observations gleaned from sensory data to compute the likelihood of performing a particular action. Other machine learning algorithms such as Deep Q-Learning, Monte Carlo Tree Search, and Temporal Difference Learning have all been shown to perform worse than BRL.

**Weather forecast:** Bayesian machine learning allows us to predict weather more accurately using Bayesian inference. Based on historical information such as temperature, humidity, and other variables, Bayes' theorem can be used to predict current weather patterns and precipitation chances. Bayesian models are superior to traditional techniques because they take into account the past behavior of the system being modeled and provide predictable probability distributions of outcomes.

**Issues in Bayesian learning methods:**

Bayesian learning requires initial knowledge of many probabilities regarding the prior probabilities.If we don't already know the prior probabilities, we can estimate them using the background information given, data that was previously available, and the underlying distribution assumptions. In Bayesian learning methods, the Baye's optimal hypothesis must be determined in the general situation with a cost that is linear in the number of candidate hypotheses. However, this computational cost can be greatly decreased in a few unique circumstances.

**Bibliography:**

# Essential Statistics, 3rd EditionBy William Navidi, Barry Monk

* <https://medium.com/@hmnathel/intro-to-bayes-theorem-8378648337fa>
* <https://wso2.com/blog/research/part-one-introduction-to-bayesian-learning/>
* <https://www.statlect.com/glossary/prior-probability>
* <https://towardsdatascience.com/bayes-rule-with-a-simple-and-practical-example-2bce3d0f4ad0?gi=1ab2fdd53b2c>