

# Bayesian Machine Learning

Eitel J.M. Lauría  
Marist College, USA

## INTRODUCTION

Bayesian methods provide a probabilistic approach to machine learning. The Bayesian framework allows us to make inferences from data using probability models for values we observe and about which we want to draw some hypotheses. Bayes theorem provides the means of calculating the probability of a hypothesis (posterior probability) based on its prior probability, the probability of the observations and the likelihood that the observational data fit the hypothesis.

$$P(H | D) = \frac{P(D | H) \cdot P(H)}{P(D)} \quad (1)$$

$P(H | D)$  is defined as the probability of a certain hypothesis based on a set of observational data given a certain context (posterior probability of hypothesis  $H$ );  $P(D | H)$  is the likelihood of the observations given a certain hypothesis;  $P(H)$  is the intrinsic probability of hypothesis  $H$ , before considering the evidence  $D$  (prior probability); and  $P(D)$  is the probability of the observations, independent of the hypothesis, that can be interpreted as a normalizing constant. Bayes rule can therefore be reformulated as shown in expression (1). This means that the probability of the hypothesis is being updated by the likelihood of the observed data.

$$P(H | D) \propto P(D | H) \cdot P(H) \quad (2)$$

## BACKGROUND

The practical application of Bayes rule to machine learning is rather straightforward. Given a set of hypotheses  $H$  and a set  $D$  of observational data we can estimate the most probable hypothesis  $H$  given  $D$ , by comparing different instances of the previous expression for each hypothesis  $H$  and choosing the one that holds the largest posterior probability (also called maximum a posteriori probability or MAP).

$$\text{Most probable } H \equiv H_{\text{MAP}} = \arg \max [P(D | H) \cdot P(H)] \quad (3)$$

Suppose we have a classification problem where the class variable is denoted by  $C$  and can take values  $c_1, c_2, \dots, c_k$ . Consider a data sample  $D$  represented by  $n$  attributes  $A_1, A_2, \dots, A_n$  of which the observations  $(a_1, a_2, \dots, a_n)$  have been taken for each instance of  $D$ . Suppose that each instance of the data sample  $D$  is classified as  $c_1, c_2, \dots, c_k$ . The Bayesian approach to classifying a new instance would then be to assign the most probable target value (a class value of type  $c_i$ ) by calculating the posterior probability for each class given the training data set, and from them choosing the one that holds the maximum a posteriori probability.

$$c_{\text{MAP}} = \arg \max_{c_i \in C} [P(D | c_i) \cdot P(c_i)] \quad (4)$$

## NAIVE BAYES CLASSIFICATION

Although the idea of applying full-blown Bayesian criteria to analyze a hypothesis space in search of the most feasible hypothesis is conceptually attractive, it usually fails to deliver in practical settings. Although we can successfully estimate  $P(c_i)$  from the training data, calculating the joint probability  $P(D | c_i)$  is usually not feasible: unless we have a very large training data set, we would end up with estimates that are representative of a small fraction of the instance space and are therefore unreliable. The naive Bayesian classifier attempts to solve this problem by making the following assumptions:

- Conditional independence among attributes of the data sample. This means that the posterior probability of  $D$ , given  $c_i$  is equal to the product of the posterior probability of each attribute.

$$P(D | c_i) = \prod_{j=1}^n P(A_j = a_j | c_i) \quad , c_i \in C \quad (5)$$

- The conditional probabilities of each individual attribute can be estimated from the frequency distributions of the sample data set  $D$  as  $N_{ij}/N_i$ , where  $N_{ij}$  is the number of training examples for which attribute  $A_j = a_j$  and class value is  $c_i$ ; and  $N_i$  is the number of training examples for which the class value is  $c_i$ . If the prior probabilities  $P(c_i)$  are not known, they can also be estimated by drawing its probabilities from the sample data set of frequency distributions.
- To solve the cases in which there are very few or no instances in the data set for which  $A_j = a_j$  given a certain class value  $c_i$ , which would in turn render poor estimates of  $P(A_j = a_j | c_i)$  or make it equal to zero, a common approach is to estimate  $P(A_j = a_j | c_i)$  as:

$$P(A_j = a_j | c_i) = \frac{N_{ij} + \alpha_{ij}}{N_i + \alpha_i} \tag{6}$$

where  $\alpha_{ij}$  and  $\alpha_i$  can be seen as fictitious counts coming out of our prior estimate of the probability we wish to determine. In rigor, this implies considering a conjugate prior probability given by a Dirichlet distribution (for more details see Ramoni & Sebastiani, 1999). A typical method for choosing  $\alpha_{ij}$  and  $\alpha_i$  in the absence of other information is to

assume uniform distribution of the counts, which means that if an attribute has  $r$  possible values,  $\alpha_{ij} = 1$  and  $\alpha_i = r$ . This results in:

$$P(A_j = a_j | c_i) = \frac{N_{ij} + 1}{N_i + r} \tag{7}$$

These assumptions have the effect of substantially reducing the number of distinct conditional probability terms that must be estimated from the training data. To illustrate the use of the naïve Bayes classifier, consider the example in Table 1 adapted from Mitchell (1997). We are dealing with records reporting on weather conditions for playing tennis. The task is to build a classifier that, by learning from previously collected data, is able to predict the chances of playing tennis based on new weather reports. We can estimate the class probabilities  $P(\text{play}=\text{yes})$  and  $P(\text{play}=\text{no})$  by calculating their frequency distributions as follows:

$$P(\text{play}=\text{yes}) = (\# \text{ of instances were play=yes}) / (\text{total} \# \text{ of instances}) = 9/14$$

$$P(\text{play}=\text{no}) = (\# \text{ of instances were play=no}) / (\text{total} \# \text{ of instances}) = 5/14$$

The conditional probabilities can be estimated by applying equation 6, as shown in Table 1(d). For a new weather report  $W = \{\text{outlook}=\text{rain}, \text{temp}=\text{hot}, \text{Humidity}=\text{high}, \text{windy}=\text{false}\}$  the classifier would compute

Table 1. Weather data set

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	No

(a) List of Instances

(play = yes)	9/14
(play = no)	5/14

(c) Prior Class Probabilities

Attribute Name	Values
outlook	sunny, overcast, rainy
temperature	hot, mild cool
humidity	high, normal
windy	true, false
play (class attribute)	yes, no

(b) List of Attributes

	play			play	
	yes	no		yes	no
Outlook			Humidity		
sunny	3/12	4/8	high	4/11	5/7
overcast	5/12	1/8	normal	7/11	2/7
rain	4/12	3/8			
Temperature			Windy		
hot	3/12	3/8	false	4/11	4/7
mild	5/12	3/8	true	7/11	3/7
cold	4/12	2/8			

(d) Conditional Probabilities

5 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

[www.igi-global.com/chapter/bayesian-machine-learning/14242](http://www.igi-global.com/chapter/bayesian-machine-learning/14242)

## Related Content

---

### **Narrating the Culture of Resistance in Harijan Community in an Urban Area of Bangladesh: An Anthropological Study**

Goutam Kumar Dutta, Md. Mostafizur Rahman, S. M. Monirul Ahasanand Md. Musfikur Rahman (2022). *International Journal of Information Systems and Social Change* (pp. 1-11).

[www.irma-international.org/article/narrating-the-culture-of-resistance-in-harijan-community-in-an-urban-area-of-bangladesh/303593](http://www.irma-international.org/article/narrating-the-culture-of-resistance-in-harijan-community-in-an-urban-area-of-bangladesh/303593)

### **Do We Really Need Elderly-Friendly Integrated Care Portals?: Mind Blocks to Their Adoption and Implementation**

Chabi Gupta (2023). *International Journal of Information Systems and Social Change* (pp. 1-6).

[www.irma-international.org/article/do-we-really-need-elderly-friendly-integrated-care-portals/314564](http://www.irma-international.org/article/do-we-really-need-elderly-friendly-integrated-care-portals/314564)

### **Inclusion of Social Subsystem Issues in IT Investment Decisions: An Empirical Assessment**

Sherry D. Ryanand Michael S. Gates (2006). *Advanced Topics in Information Resources Management, Volume 5* (pp. 164-183).

[www.irma-international.org/chapter/inclusion-social-subsystem-issues-investment/4647](http://www.irma-international.org/chapter/inclusion-social-subsystem-issues-investment/4647)

### **From a FabLab towards a Social Entrepreneurship and Business Lab**

Alicia Guerra Guerraand Lyda Sánchez de Gómez (2016). *Journal of Cases on Information Technology* (pp. 1-21).

[www.irma-international.org/article/from-a-fablab-towards-a-social-entrepreneurship-and-business-lab/173721](http://www.irma-international.org/article/from-a-fablab-towards-a-social-entrepreneurship-and-business-lab/173721)

### **Integrating the Fragmented Pieces of IS Research Paradigms and Frameworks: A Systems Approach**

Manuel Mora, Ovsei Gelman, Guisseppi Forgionne, Doncho Petkovand Jeimy Cano (2009). *Best Practices and Conceptual Innovations in Information Resources Management: Utilizing Technologies to Enable Global Progressions* (pp. 1-22).

[www.irma-international.org/chapter/integrating-fragmented-pieces-research-paradigms/5508](http://www.irma-international.org/chapter/integrating-fragmented-pieces-research-paradigms/5508)