From Bayesian Inference to Logical Bayesian Inference

Chenguang Lu

From Bayesian Inference to Logical Bayesian Inference:

A New Mathematical Frame for Semantic Communication and Machine Learning

Chenguang Lu [0000-0002-8669-0094]

College of Intelligence Engineering and Mathematics,
Liaoning Engineering and Technology University, Fuxin, Liaoning, 123000, China
lcguang@foxmail.com

Abstract. Bayesian Inference (BI) uses the Bayes' posterior whereas Logical Bayesian Inference (LBI) uses the truth function or membership function as the inference tool. LBI was proposed because BI was not compatible with the classical Bayes' prediction and didn't use logical probability and hence couldn't express semantic meaning. In LBI, statistical probability and logical probability are strictly distinguished, used at the same time, and linked by the third kind of Bayes' Theorem. The Shannon channel consists of a set of transition probability functions whereas the semantic channel consists of a set of truth functions. When a sample is large enough, we can directly derive the semantic channel from Shannon's channel. Otherwise, we can use parameters to construct truth functions and use the Maximum Semantic Information (MSI) criterion to optimize the truth functions. The MSI criterion is equivalent to the Maximum Likelihood (ML) criterion, and compatible with the Regularized Least Square (RLS) criterion. By matching the two channels one with another, we can obtain the Channels' Matching (CM) algorithm. This algorithm can improve multi-label classifications, maximum likelihood estimations (including unseen instance classifications), and mixture models. In comparison with BI, LBI 1) uses the prior $P(X)$ of X instead of that of Y or θ and fits cases where the source $P(X)$ changes, 2) can be used to solve the denotations of labels, and 3) is more compatible with the classical Bayes' prediction and likelihood method. LBI also provides a confirmation measure between -1 and 1 for induction.

Keywords: Bayes' Theorem, Bayesian inference, MLE, MAP, Semantic information, Machine learning, Confirmation measure.

1 Introduction[1]

Bayesian Inference (BI) [1, 2] was proposed by Bayesians. Bayesianism and Frequentism are contrary [3]. Frequentism claims that probability is objective and can be

[1] This paper is a condensed version in English. The original Chinese version: http://survivor99.com/lcg/CM/Homepage-NewFrame.pdf

defined as the limit of the relative frequency of an event; whereas Bayesianism claims that probability is subjective or logical. Some Bayesians consider probability as degree of belief [3] whereas others, such as Keynes [4], Carnap [5], and Jaynes [6], so-called logical Bayesians, consider probability as the truth value. There are also minor logical Bayesians, such as Reichenbach [7] as well as the author of this paper, who use frequency to explain the logical probability and truth function.

Many frequentists, such as Fisher [8] and Shannon [9], also use Bayes' Theorem, but they are not Bayesians. Frequentist main tool for hypothesis-testing is Likelihood Inference (LI), which has achieved great successes. However, LI cannot make use of prior knowledge. For example, after the prior distribution $P(x)$ of an instance x is changed, the likelihood function $P(x|\theta_j)$ will be no longer valid[1]. To make use of prior knowledge and to emphasize subjective probability, some Bayesians proposed BI [1] which used the Bayesian posterior $P(\theta|\mathbf{X})$, where $\mathbf{X}$ was a sequence of instances, as the inference tool. The Maximum Likelihood Estimation (MLE) was revised into the Maximum A Posterior estimation (MAP) [2]. Demonstrating some advantages especially for working with small samples and for solving the frequency distribution of a frequency producer, BI also has some limitations. The main limitations are: 1) It is incompatible with the classical Bayes' prediction as shown by Eq. (1); 2) It does not use logical probabilities or truth functions and hence cannot solve semantic problems. To overcome these limitations, we propose Logical Bayesian Inference (LBI), following earlier logical Bayesians to use the truth function as the inference tool and following Fisher to use the likelihood method. The author also set up new mathematical frame employing LBI to improve semantic communication and machine learning.

LBI has the following features:

- It strictly distinguishes statistical probability and logical probability, uses both at the same time, and links both by the third kind of Bayes' Theorem, with which the likelihood function and the truth function can be converted from one to another.
- It also uses frequency to explain the truth function, as Reichenbach did, so that optimized truth function can be used as transition probability function $P(y_j|x)$ to make Bayes' prediction even if $P(x)$ is changed.
- It brings truth functions and likelihood functions into information formulas to obtain the generalized Kullback-Leibler (KL) information and the semantic mutual information. It uses the Maximum Semantic Information (MSI) criterion to optimize truth functions. The MSI criterion is equivalent to the Maximum Likelihood (ML) criterion and compatible with the Regularized Least Squares (RLS) criterion [10].

Within the new frame, we convert sampling sequences into sampling distributions and then use the cross-entropy method [10]. This method has become popular in recent two decades because it is suitable to larger samples and similar to information theoretical method. This study is based on the author's studies twenty years ago on semantic information theory with the cross-entropy and mutual cross-entropy as tools [11-14]. This study also relates to the author's recent studies on machine learning for simplifying multi-label classifications [15], speeding the MLE for tests and unseen instance classifications [16], and improving the convergence of mixture models [17].

In the following sections, the author will discuss why LBI is employed (Section 2), introduce the mathematical basis (Section 3), state LBI (Section 4), introduce its applications to machine learning (Section 5), discuss induction (Section 6), and summarize the paper finally.

2 From Bayes' Prediction to Logical Bayesian Inference

Definition 1:

- x: an instance or data point; X: a random variable; $X=x\in U=\{x_1, x_2, \ldots, x_m\}$.
- y: a hypothesis or label; Y: a random variable; $Y=y\in V=\{y_1, y_2, \ldots, y_n\}$.
- θ: a model or a set of model parameters. For given y_j, $\theta=\theta_j$.
- $\mathbf{X}_j$: a sample or sequence of data point $x(1), x(2), \ldots, x(N_j)\in U$. The data points come from Independent and Identically Distributed (IID) random variables.
- **D:** a sample or sequence of examples $\{(x(t), y(t))|t=1 \text{ to } N;\ x(t)\in U; y(t)\in V\}$, which includes n different sub-samples $\mathbf{X}_j$. If **D** is large enough, we can obtain distribution $P(x, \mathrm{y})$ from **D**, and distribution $P(x|y_j)$ from $\mathbf{X}_j$.

A Shannon's channel $P(y|x)$ consists of a set of Transition Probability Functions (TPF) $P(y_j|x)$, $j=1, 2, \ldots$ A TPF $P(y_j|x)$ is a good prediction tool. With the Bayes' Theorem II (discussed in Section 3), we can make probability prediction $P(x|y_j)$ according to $P(y_j|x)$ and $P(x)$. Even if $P(x)$ changes into $P'(x)$, we can still obtain $P'(x|y_j)$ by

$$P'(x \mid y_j) = P(y_j \mid x)P'(x) / \sum_i P(y_j \mid x_i)P'(x_i) \tag{1}$$

We call this probability prediction as "classical Bayes' prediction". However, if samples are not large enough, we cannot obtain continuous distributions $P(y_j|x)$ or $P(x|y_j)$. Therefore, Fisher proposed Likelihood Inference (LI) [7].

For given $\mathbf{X}_j$, the likelihood of θ_j is

$$P(\mathbf{X}_j \mid \theta_j) = P(x(1), x(2), ..., x(N) | \theta_j) = \prod_{t=1}^{N} P(x(t) \mid \theta_j) \tag{2}$$

We use θ_j instead of θ in the above equation because unlike the model θ in BI, the model in LI does not have a probability distribution. If there are N_{ji} x_i in $\mathbf{X}_j$, then $P(x_i|y_j)=N_{ji}/N_j$, and the likelihood can be expressed by a negative cross entropy:

$$\begin{aligned} &\log P(\mathbf{X}_j \mid \theta_j) = \log \prod_i P(x_i \mid \theta_j)^{N_{ji}} \\ &= N_j \sum_i P(x_i \mid y_j) \log P(x_i|\theta_j) = -N_j H(X \mid \theta_j) \end{aligned} \tag{3}$$

For conditional sample $\mathbf{X}_j$ whose distribution is $P(x|j)$ (the label is uncertain), we can find the MLE:

$$\theta_j^* = \arg\max_{\theta_j} P(\mathbf{X}_j \mid \theta_j) = \arg\max_{\theta_j} \sum_i P(x_i \mid j) \log P(x_i \mid \theta_j) \tag{4}$$

When $P(x|\theta_j)=P(x|j)$, $H(X|\theta_j)$ reaches its minimum.

The main limitation of LI is that it cannot make use of prior knowledge, such as $P(x)$, $P(y)$, or $P(\theta)$, and does not fit cases where $P(x)$ may change. BI brings the prior distribution $P(\theta)$ of θ into the Bayes' Theorem II to have [2]

$$P(\theta \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \theta) P(\theta)}{P_\theta(\mathbf{X})}, \quad P_\theta(\mathbf{X}) = \sum_j P(\mathbf{X} \mid \theta_j) P(\theta_j) \tag{5}$$

where $P_\theta(\mathbf{X})$ is the normalized constant related to θ. For one Bayesian posterior, we need n or more likelihood functions. The MLE becomes the MAP:

$$\theta_j^* = \arg\max_{\theta_j} P(\theta \mid \mathbf{X}_j) = \arg\max_{\theta_j} [\sum_i P(x_i \mid j) \log P(x_i \mid \theta_j) + \log P(\theta_j)] \tag{6}$$

where $P_\theta(\mathbf{X})$ is neglected. It is easy to find that 1) if $P(\theta)$ is neglected or is an equiprobable distribution, the MAP is equivalent to the MLE; 2) while the sample's size N increases, the MAP gradually approaches the MLE.

There is also the Bayesian posterior of Y:

$$P(Y \mid \mathbf{X}, \theta) = \frac{P(\mathbf{X} \mid \theta) P(Y)}{P_\theta(\mathbf{X})}, \quad P_\theta(\mathbf{X}) = \sum_j P(\mathbf{X} \mid \theta_j) P(y_j) \tag{7}$$

It is different from $P(\theta|\mathbf{X})$ because $P(Y|\mathbf{X},\theta)$ is a distribution over the label space whereas $P(\theta|\mathbf{X})$ is a distribution over the parameter space. The parameter space is larger than the label space. $P(Y|\mathbf{X}, \theta)$ is easier understood than $P(\theta|X)$. $P(Y|\mathbf{X}, \theta)$ is also often used, such as for mixture models and hypothesis-testing.

BI has some advantages: 1) It considers the prior of Y or θ so that when $P(X)$ is unknown, $P(Y)$ or $P(\theta)$ is also useful, especially for small samples. 2) It can convert the current posterior $P(\theta|\boldsymbol{X})$ into the next prior $P(\theta)$. 3) The distribution $P(\theta|\boldsymbol{X})$ over θ space will gradually concentrate as the sample's size N increases. When $N \to \infty$, only $P(\theta^*|\mathbf{X})=1$, where θ^* is the MAP. So, $P(\theta|X)$ can intuitively show learning results.

However, there are also some serious problems with BI:

1) About Bayesian prediction

BI predicts the posterior and prior distributions of x by [2]:

$$P_\theta(x \mid \mathbf{X}) = \sum_j P(x|\theta_j) P(\theta_j \mid \mathbf{X}) \text{ and } P_\theta(x) = \sum_j P(x|\theta_j) P(\theta_j) \tag{8}$$

From a huge sample $\mathbf{D}$, we can directly obtain $P(x|y_j)$ and $P(x)$. However, BI cannot ensure $P_\theta(x|\mathbf{X})=P(x|y_j)$ or $P_\theta(x)=P(x)$. After $P(x)$ changes into $P'(x)$, BI cannot obtain the posterior that is equal to $P'(x|y_j)$ in Eq. (1). Hence, the Bayesian prediction is not compatible with the classical Bayes' prediction. Therefore, we need an inference tool that is like the TPF $P(y_j|x)$ and is constructed with parameters.

2) About logical probability

BI does not use logical probability because logical probability is not normalized; nevertheless, all probabilities BI uses are normalized. Consider labels "Non-rain", "Rain", "Light rain", "Moderate rain", "Light to moderate rain", …, in a weather forecast. The sum of their logical probabilities is greater than 1. The conditional logical probability or truth value of a label with the maximum 1 is also not normalized. BI uses neither truth values nor truth functions and hence cannot solve the denotation (or semantic meaning) of a label. Fuzzy mathematics [18, 19] uses membership functions, which can also be used as truth functions. Therefore, we need an inference method that can derive truth functions or membership functions from sampling distributions.

3) About prior knowledge

In BI, $P(\theta)$ is subjective. However, we often need objective prior knowledge. For example, to make probability prediction about a disease according to a medical test result "positive" or "negative", we need to know the prior distribution $P(x)$ [16]. To predict a car's real position according to a GPS indicator on a GPS map, we need to know the road conditions, which tell $P(x)$[1].

4) About optimization criterion

According to Popper's theory [20], a hypothesis with less LP can convey more information. Shannon's information theory [9] contains a similar conclusion. The MAP is not well compatible with the information criterion.

The following example can further explain why we need LBI.

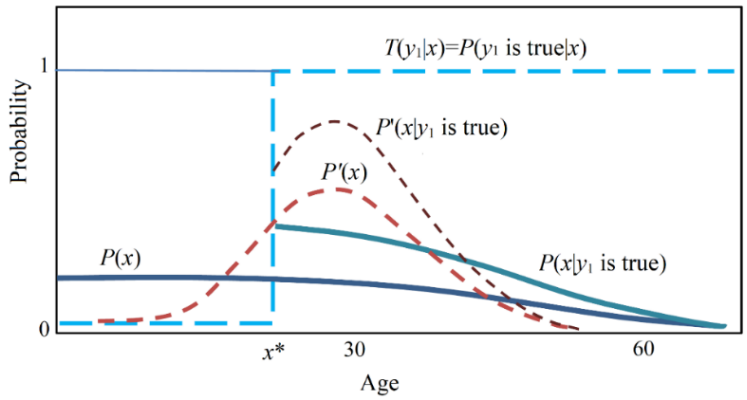

Fig. 1. Solving the denotation of y_1="x is adult" and probability prediction $P'(x|y_1$ is true).

Example 1. Given the age population prior distribution $P(x)$ and posterior distribution $P(x|$ "adult" is true), which are continuous, please answer:

1) How do we obtain the denotation (e. g. the truth function) of "adult" (see Fig. 1)?

2) Can we make a new probability prediction or produce new likelihood function with the denotation when $P(x)$ is changed into $P'(x)$?

3) If the set {Adult} is fuzzy, can we obtain its membership function?

It is difficult to answer these questions using either LI or BI. Nevertheless, using LBI, we can easily obtain the denotation and make the new probability prediction.

3 Mathematical Basis: Three Kinds of Probabilities and Three Kinds of Bayes' Theorems

All probabilities [3] can be divided into three kinds:

1) Statistical probability: relative frequency or its limit of an event;
2) Logical Probability (LP): how frequently a hypothesis is judged true or how true a hypothesis is;
3) Predicted (or subjective) probability: possibility or likelihood.

We may treat the predicted probability as the hybrid of the former two kinds. Hence, there are only two kinds of basic probabilities: the statistical and the logical.

A hypothesis or label has both statistical (or selected) probability and LP. They are very different. Consider labels in a weather forecast: "Light rain" and "Light to heavy rain". The former has larger selected probability and less LP. The LP of a tautology, such as "He is old or not old", is 1 whereas its selected probability is close to 0.

Each of existing probability systems [3, 5-7] only contains one kind of probabilities. Now we define a probability system with both statistical probabilities and LPs.

Definition 2 A label y_j is also a predicate $y_j(X)$= "$X \in A_j$." For y_j, U has a subset A_j, every x in which makes y_j true. Let $P(Y=y_j)$ denote the statistical probability of y_j, and $P(X \in A_j)$ denote the LP of y_j. For simplicity, let $P(y_j)=P(Y=y_j)$ and $T(A_j)=P(X \in A_j)$.

We call $P(X \in A_j)$ the LP because according to Tarski's theory of truth [21], $P(X \in A_j)=P$("$X \in A_j$" is true)$=P(y_j$ is true). Hence, the conditional LP $T(A_j|X)$ of y_j for given X is the feature function of A_j and the truth function of y_j. Hence the LP of y_j is

$$T(A_j) = \sum_i P(x_i) T(A_j \mid x_i) \tag{9}$$

According to Davidson's truth-conditional semantics [21], $T(A_j|X)$ ascertains the semantic meaning of y_j. Note that statistical probability distributions, such as $P(Y)$, $P(Y|x_i)$, $P(X)$, and $P(X|y_j)$, are normalized; however, LP distributions are not normalized. In general, $T(A_1)+T(A_2)+\ldots+T(A_n)>1$; $T(A_1|x_i)+T(A_2|x_i)+\ldots+T(A_n|x_i)>1$.

If A_j is fuzzy, $T(A_j|X)$ becomes the membership function, and $T(A_j)$ becomes the fuzzy event probability defined by Zadeh [19]. For fuzzy sets, we use θ_j to replace A_j. Then $T(\theta_j|X)$ becomes the membership function of θ_j. That means

$$m_{\theta_j}(X) = T(\theta_j \mid X) = T(y_j \mid X) \tag{10}$$

We can also treat θ_j as a sub-model of a predictive model θ. In this paper, the likelihood function $P(X|\theta_j)$ is equal to $P(X|y_j; \theta)$ in popular likelihood method. $T(\theta_j|X)$ is different from $P(\theta|X)$ and is longitudinally normalized, e. g.,

$$\max(T(\theta_j|X))=\max(T(\theta_j|x_1), T(\theta_j|x_2), \ldots, T(\theta_j|x_m)) =1 \tag{11}$$

There are three kinds of Bayes' Theorems, which are used by Bayes [23], Shannon [9], and the author respectively.

Bayes' Theorem I (used by Bayes): Let sets $A, B \in 2^U$, A^c be the complementary set of A, $T(A)=P(X \in A)$, and $T(B)= P(X \in B)$. Then

$$T(B|A)=T(A|B)T(B)/T(A),\ T(A)= T(A|B)T(B)+ T(A|B^c)T(B^c) \tag{12}$$

There is also a symmetrical formula for $T(A|B)$. Note there are only one random variable X and two logical probabilities $T(A)$ and $T(B)$.

Bayes' Theorem II (used by Shannon):

$$P(X \mid y_j) = P(y_j \mid X)P(X) / P(y_j),\ P(y_j) = \sum_i P(x_i)P(y_j \mid x_i) \tag{13}$$

There is also a symmetrical formula for $P(y_j|X)$ or $P(Y|x_i)$. Note there are two random variables and two statistical probabilities.

Bayes' Theorem III:

$$P(X \mid \theta_j) = T(\theta_j \mid X)P(X) / T(\theta_j),\ T(\theta_j) = \sum_i P(x_i)T(\theta_j \mid x_i) \tag{14}$$

$$T(\theta_j \mid X)=P(X \mid \theta_j)T(\theta_j) / P(X),\ T(\theta_j) = 1 / \max(P(X \mid \theta_j) / P(X)) \tag{15}$$

The two formulas are asymmetrical because there is a statistical probability and a logical probability. $T(\theta_j)$ in Eq. (15) may be called longitudinally normalizing constant.

The Proof of Bayes' Theorem III: The joint probability $P(X, \theta_j)=P(X=x, X \in \theta_j)$, then $P(X|\theta_j)T(\theta_j)=P(X=x, X \in \theta_j)=T(\theta_j|X)P(X)$. Hence there is

$$P(X \mid \theta_j) = P(X)T(\theta_j \mid X) / T(\theta_j),\ T(\theta_j|X) = T(\theta_j)P(X \mid \theta_j) / P(X)$$

Since $P(X|\theta_j)$ is horizontally normalized, $T(\theta_j)=\sum_i P(x_i)T(\theta_j|x_i)$. Since $T(\theta_j|X)$ is longitudinally normalized, there is

$$1= \max[T(\theta_j)P(X|\theta_j)/P(X)]=T(\theta_j)\max[P(X|\theta_j)/P(X)]$$

Hence $T(\theta_j)=1/\max[P(X|\theta_j)/P(X)]$. **QED.**

Eq. (15) can also be expressed as

$$T(\theta_j \mid X)=[P(X \mid \theta_j) / P(X)] / \max[P(X \mid \theta_j) / P(X)] \tag{16}$$

Using this formula, we can answer questions of **Example 1** in Section 2 to obtain the denotation of "Adult" and the posterior distribution $P'(x|y_1$ is true) as shown in Fig. 1.

4 Logical Bayesian Inference (LBI)

LBI has three tasks:

1) To derive truth functions or a semantic channel from **D** or sampling distributions (e.g., multi-label learning [24, 25]);

2) To select hypotheses or labels to convey information for given x or $P(X|j)$ according to the semantic channel (e. g., multi-label classification);

3) To make logical Bayes' prediction $P(X|\theta_j)$ according to $T(\theta_j|X)$ and $P(X)$ or $P'(X)$.

The third task is simple. We can use Eq. (14) for this task.

For the first task, we first consider continuous sampling distributions from which we can obtain The Shannon channel $P(Y|X)$. The TPF $P(y_j|X)$ has an important property: $P(y_j|X)$ by a constant k can make the same probability prediction because

$$\frac{P'(X)kP(y_j \mid X)}{\sum_i P'(x_i)kP(y_j \mid x_i)} = \frac{P'(X)P(y_j \mid X)}{\sum_i P'(x_i)P(y_j \mid x_i)} = P'(X \mid y_j) \tag{17}$$

A semantic channel $T(\theta|X)$ consists of a set of truth functions $T(\theta_j|X)$, j=1, 2, …, n. According to Eq. (17), if $T(\theta_j|X) \propto P(y_j|X)$, then $P(X|\theta_j)=P(X|y_j)$. Hence the optimized truth function is

$$T^*(\theta_j|X)=P(y_j|X)/\max(P(y_j|X)) \tag{18}$$

We can prove[1] that the truth function derived from (18) is the same as that from Wang's random sets falling shadow theory [26]. According to the Bayes' Theorem II, from Eq. (18), we obtain

$$T^*(\theta_j|X)= [P(X|y_j)/P(X)]/\max[P(X|y_j)/P(X)] \tag{19}$$

Eq. (19) is more useful in general because it is often hard to find $P(y_j|X)$ or $P(y_j)$ for Eq. (18). Eqs. (18) and (19) fit cases involving large samples. When samples are not large enough, we need to construct truth functions with parameters and to optimize them.

The semantic information conveyed by y_j about x_i is defined with log-normalized-likelihood [12, 14]:

$$I(x_i;\theta_j) = \log\frac{P(x_i \mid \theta_j)}{P(x_i)} = \log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{20}$$

For an unbiased estimation y_j, its truth function may be expressed by a Gaussian distribution without the coefficient: $T(\theta_j|X)=\exp[-(X-x_j)^2/(2d^2)]$. Hence

$$I(x_i; \theta_j)=\log[1/T(\theta_j)]-(X-x_j)^2/(2d^2) \tag{21}$$

The $\log[1/T(\theta_j)]$ is the Bar-Hillel-Carnap semantic information measure [27]. Eq. (21) shows that the larger the deviation is, the less information there is; the less the LP is, the more information there is; and, a wrong estimation may convey negative information. These conclusions accord with Popper's thought [20].

To average $I(x_i; \theta_j)$, we have

$$I(X;\theta_j) = \sum_i P(x_i \mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)} = \sum_i P(x_i \mid y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{22}$$

where $P(x_i|y_j)$ (i=1, 2, …) is the sampling distribution, which may be unsmooth or discontinuous. Hence, the optimized truth function is

$$T^*(\theta_j \mid X) = \arg\max_{T(\theta_j|X)} I(X;\theta_j) = \arg\max_{T(\theta_j|X)} \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \quad (23)$$

It is easy to prove that when $P(X|\theta_j)=P(X|y_j)$ or $T(\theta_j|X) \propto P(y_j|X)$, $I(X;\theta_j)$ reaches its maximum and is equal to the KL information. If we only know $P(X|y_j)$ without knowing $P(X)$, we may assume that X is equiprobable to obtain the truth function.

To average $I(X; \theta_j)$ for different Y, we have [12, 14]

$$\begin{aligned} &I(X;\theta) = H(\theta) - H(\theta \mid X) \\ &H(\theta) = -\sum_j P(y_j) \log T(\theta_j), \;\; H(\theta \mid X) = \sum_j \sum_i P(x_i, y_j)(x_i - x_j)^2 / \left(2d_j^{\,2}\right) \end{aligned} \quad (24)$$

Clearly, the MSI criterion is like the RLS criterion. $H(\theta|X)$ is like the mean squared error, and $H(\theta)$ is like the negative regularization term. The relationship between the log normalized likelihood and generalized KL information is

$$\log \prod_i \left[\frac{P(x_i|\theta_j)}{P(x_i)} \right]^{N_{ji}} = N_j \sum_i P(x_i \mid y_j) \log \frac{P(x_i|\theta_j)}{P(x_i)} = N_j I(X;\theta_j) \quad (25)$$

The MSI criterion is equivalent to the ML criterion because $P(X)$ does not change when we optimize θ_j.

For the second task of LBI, given x_i, we select a hypothesis or label by the classifier

$$y_j^* = h(x) = \arg\max_{y_j} \log I(\theta_j; x) = \arg\max_{y_j} \log \frac{T(\theta_j \mid x)}{T(\theta_j)} \quad (26)$$

This classifier produces a noiseless Shannon's channel. Using $T(\theta_j)$, we can overcome the class-imbalance problem [24]. If $T(\theta_j|x) \in \{0,1\}$, the classifier becomes

$$y_j^* = h(x) = \arg\max_{y_j \text{ with } T(A_j|x)=1} \log[1/T(A_j)] = \arg\min_{y_j \text{ with } T(A_j|x)=1} T(A_j) \quad (27)$$

It means that we should select a label with the least LP and hence with the richest connotation. The above method of multi-label learning and classification is like the Binary Relevance (BR) method [25]. However, the above method does not demand too much of samples and can fit cases where $P(X)$ changes (See [15] for details).

5 Logical Bayesian Inference for Machine Learning

In Section 4, we have introduced the main method of using LBI for multi-label learning and classification. From LBI, we can also obtain an iterative algorithm, the Channels' Matching (CM) algorithm, for the MLE [16] and mixture models [17].

For unseen instance classifications, we assume that observed condition is $Z \in C=\{z_1, z_2, \ldots\}$; the classifier is $Y=f(Z)$; a true class or true label is $X \in U=\{x_1, x_2, \ldots\}$; a sample is $\mathbf{D}=\{(x(t); z(t))|t=1, 2, \ldots, N; x(t) \in U; z(t) \in C\}$. From **D**, we can obtain $P(X, Z)$. The iterative process is as follows.

Step I (the semantic channel matches the Shannon channel): For a given classifier $Y=f(Z)$, we obtain $P(Y|X)$, $T(\theta|X)$, and conditional information for given Z

$$I(X;\theta_j|Z) = \sum_i P(X_i \mid Z) \log \frac{T(\theta_j \mid X_i)}{T(\theta_j)},\ j=1, 2, \ldots, n \tag{28}$$

Step II (the Shannon channel matches the semantic channel): The classifier is

$$y_j = f(Z) = \arg\max_{y_j} I(X;\theta_j \mid Z),\quad j=1, 2, \ldots, n \tag{29}$$

Repeating the above two steps, we can achieve the MSI and ML classification. The convergence can be proved with the help of $R(G)$ function [16].

For mixture models, the aim of **Step II** is to minimize the Shannon mutual information R minus the semantic mutual information G [17]. The convergence of the CM for mixture models is more reliable than that of the EM algorithm[2].

6 Confirmation Measure b^* for Induction

Early logical Bayesians [4-7] were also inductivists who used the conditional LP or truth function to indicate the degree of inductive support. However, contemporary inductivists use the confirmation measure or the degree of belief between -1 and 1 for induction [28, 29]. By LBI, we can derive a new confirmation measure $b^* \in [-1,1]$.

Now we use the medical test as an example to introduce the confirmation measure b^*. Let x_1 be a person with a disease, x_0 be a person without the disease, y_1 be the test-positive, and y_0 be the test-negative. The y_1 also means a universal hypothesis "For all people, if one's testing result is positive, then he/she has the disease". According to Eq. (18), the truth value of proposition $y_1(x_0)$ (x_0 is the counterexample of y_1) is

$$b'^* = T^*(\theta_1|x_0) = P(y_1|x_0)/P(y_1|x_1) \tag{30}$$

We define the confirmation measure b^* of y_1 by $b'^*=1-|b^*|$. The b^* can also be regarded as the optimized degree of belief of y_1. For this measure, having fewer counterexamples or $P(y_1|x_0)$ is more important than having more positive examples or $P(y_1|x_1)$. Therefore, this measure is compatible with Popper's falsification theory [30].

With the TPH $P(y_1|X)$, we can use the likelihood method to obtain the Confidence Level (CL) of y_1, which reflects the degree of inductive support of y_1. Using $T^*(\theta_1|X)$, we can obtain the same CL. And, the b^* is related to CL by[1]

[2] For the strict convergence proof, see http://survivor99.com/lcg/CM/CM4MM.html

$$b^* = \begin{cases} 1 - \mathrm{CL}'/\mathrm{CL}, & \text{if } \mathrm{CL} > 0.5 \\ \mathrm{CL}'/\mathrm{CL} - 1, & \text{if } \mathrm{CL} \le 0.5 \end{cases} \tag{31}$$

where CL'=1-CL. If the evidence or sample fully supports a hypothesis, then CL=1 and b^*=1. If the evidence is irrelevant to a hypothesis, CL=0.5 and b^*=0. If the evidence fully supports the negative hypothesis, CL=0 and b^*=-1. The b^* can indicate the degree of inductive support better than CL because inductive support may be negative. For example, the confirmation measure of "All ravens are white" should be negative.

If $|U|>2$ and $P(y_j|X)$ is a distribution over U, we may use the confidence interval to convert a predicate into a universal hypothesis, and then, to calculate its confidence level and the confirmation measure[1].

BI provides the credible level with given credible interval [3] for a parameter distribution instead of a hypothesis. The credible level or the Bayesian posterior does not well indicate the degree of inductive support of a hypothesis. In comparison with BI, LBI should be a better tool for induction.

7 Summary

This paper proposes the Logical Bayesian Inference (LBI), which uses the truth function as the inference tool like logical Bayesians and uses the likelihood method as frequentists. LBI also use frequencies to explain logical probabilities and truth functions, and hence is the combination of extreme frequentism and extreme Bayesianism. The truth function LBI uses can indicate the semantic meaning of a hypothesis or label and can be used for probability prediction that is compatible with the classical Baye's prediction. LBI is based on the third kind of Bayes' Theorem and the semantic information method. They all together form a new mathematical frame for semantic communication, machine learning, and induction. This new frame may support and improve many existing methods, such as likelihood method and fuzzy mathematics method, rather than replace them. As a new theory, it must be imperfect. The author welcomes researchers to criticize or improve it.

References

1. Fienberg, S. E.: When Did Bayesian Inference Become "Bayesian"? Bayesian Analysis 1(1), 1-40 (2006).
2. Anon: Bayesian Inference, Wikipedia: the Free Encyclopedia. https://en.wikipedia.org/wiki/Bayesian_probability, last accessed 2018/7/20.
3. Hájek, A.: Interpretations of Probability, In: Zalta, E. N. (ed.) The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), https://plato.stanford.edu/archives/win2012/entries/probability-interpret/, last accessed 2018/7/27.
4. Keynes, I. M.: A Treaties on Probability, Macmillan, London (1921).
5. Carnap, R.: Logical Fundations of Probability, The Univ. of Chicago Press, Chicago (1962).
6. Jaynes, E. T.: Probability Theory: The logic of Science, Edited by Larry Bretthorst. Cambridge University press, New York (2003).

7. Reichenbach, H.: The Theory of Probability, Univ. of California Press, Berkeley (1949).
8. Fisher, R. A.: On the mathematical foundations of theoretical statistics. Philo. Trans. Roy. Soc. A222, 309-368 (1922).
9. Shannon, C. E.: A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–429 and 623–656 (1948).
10. Goodfellow, I., Bengio, Y.: Deep Learning, The MIP Press, Cambridge (2016).
11. Lu, C.: B-fuzzy quasi-Boolean algebra and generalized mutual entropy formula. Fuzzy Systems and Mathematics (in Chinese) 5(1), 76-80 (1991).
12. Lu, C.: A Generalized Information Theory (in Chinese). China Science and Technology University Press, Hefei (1993).
13. Lu, C.: Meanings of Generalized Entropy and Generalized Mutual Information for Coding (in Chinese). J. of China Institute of Communication 15(6), 37-44 (1994).
14. Lu, C.: A Generalization of Shannon's Information Theory. Int. J. of General Systems 28(6): 453-490 (1999).
15. Lu, C.: Semantic channel and Shannon channel mutually match for multi-label classifications. ICIS2018, Beijing, China (2018).
16. Lu C.: Semantic Channel and Shannon Channel Mutually Match and Iterate for Tests and Estimations with Maximum Mutual Information and Maximum Likelihood. In: 2018 IEEE International Conference on Big Data and Smart Computing, pp. 227-234, IEEE Conference Publishing Services, Piscataway (2018).
17. Lu, C.: Channels' matching algorithm for mixture models. In: Shi et al. (Eds.) IFIP International Federation for Information Processing, pp. 321–332. Springer International Publishing, Switzerland (2017).
18. Zadeh, L. A.: Fuzzy Sets. Information and Control 8(3), 338–53 (1965).
19. Zadeh, L., A.: Probability measures of fuzzy events, J. of Mathematical, Analysis and Applications 23(2), 421-427 (1986).
20. Popper, K.: Conjectures and Refutations. Repr. Routledge, London and New York (1963/2005).
21. Tarski, A.: The semantic conception of truth: *and* the foundations of semantics. Philosophy and Phenomenological Research 4(3), 341–376 (1944).
22. Davidson, D.: Truth and meaning. Synthese 17(1), 304-323 (1967).
23. Bayes T, Price R. An essay towards solving a problem in the doctrine of chance. Philosophical Transactions of the Royal Society of London 53(0), 370–418 (1763).
24. Zhang, M. L., Zhou, Z. H.: A review on multi-label learning algorithm. IEEE Transactions on Knowledge and Data Engineering 26(8), 1819-1837(2014).
25. Zhang, M. L., Li, Y. K., Liu, X. Y., Geng, X.:Binary relevance for multi-label learning: an overview. Frontiers of Computer Science 12(2), 191-202 (2018).
26. Wang, P. Z.: From the Fuzzy Statistics to the Falling Random Subsets. in: Wang, P. P. (ed.), Advances in Fuzzy Sets, Possibility Theory and Applications, pp. 81–96. Plenum Press, New York (1983).
27. Bar-Hillel Y, Carnap R.: An outline of a theory of semantic information. Tech. Rep. No. 247, Research Lab. of Electronics, MIT (1952).
28. Hawthorne, J.: Inductive logic. In: Edward N. Zalta (ed.) The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/logic-inductive/, last accessed 2018/7/22.
29. Tentori, K., Crupi, V., Bonini, N., Osherson, D.: Comparison of confirmation measures. Cognition 103(1), 107–119 (2017).
30. Lu, C.: Semantic Information Measure with Two Types of Probability for Falsification and Confirmation, https://arxiv.org/abs/1609.07827, last accessed 2018/7/27.