The Third Kind of Bayes' Theorem Links Membership Functions to Likelihood Functions and Sampling Distributions

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. For given age population prior distribution $P(x)$ and the posterior distribution $P(x|\text{adult})$, how do we obtain the denotation of a label y="adult"? With the denotation, e.g., the membership function of class {Adult}, we can make new probability prediction, e.g., likelihood function, for changed $P(x)$. However, existing methods including Likelihood Method and Bayesian Inference cannot resolve this problem. For this purpose, the author proposes and proves the third kind of Bayes' Theorem, which includes two asymmetrical Bayes' formulas. The membership function so obtained is equivalent to that from random set statistics proposed by Peizhuang Wang. When samples are very big so that there are continuous sampling distributions $P(x, y)$, we can directly derive a group of membership functions by a new Bayes formula. If samples are not big enough, we can use the semantic information formula, a generalized Kullback-Leibler formula, to optimize the membership functions by sampling distributions. The semantic information criterion is compatible with maximum likelihood criterion and Regularized Least Squares (RLS) criterion. In comparison with the likelihood function and the Bayesian posterior, the membership function so obtained as the predictive model can be used with new source $P(x)$ to produce new likelihood function for better generalization performance. New Bayes' formulas and the semantic information method can be applied to machine learning. The paper simply introduces their applications to 1) multi-label classifications; 2) maximum mutual information classifications for unseen instances; and 3) mixture models. It is shown that new methods are very simple and reliable. It seems that the membership function so obtained can bridge the gap between logic and probability. With this membership function, we can develop a new mathematical tool: Logical Bayesian Inference.

Keywords: Bayes' theorem, Membership function, Likelihood function, Semantic information, Machine learning, Multi-label classification, Natural language processing, Logical probability.

1 Introduction

The main task of machine learning is classification. The membership function proposed by Zadeh [1] indicates the membership relation of different instances to a fuzzy class and hence should be a good tool for machine learning. The relationship between statistical probabilities and membership functions has been discussed for a long time [2-6]. Peizhuang Wang [2] explained the membership function with random set falling shadow. Thomas and others [3, 4] proposed a Bayes' formula to produce a likelihood function from a membership function and an instance prior distribution (e.g., a source). The author used Wang's random set falling shadow theory to derived the above Bayes' formula and used it to set up a semantic information theory [7-9]. However, existing methods still cannot obtain a membership function, which is compatible with random set statistics, from a likelihood function or a sampling distribution directly.

We use an example to explain this problem. Assume there is age population prior distribution $P(x)$ and the posterior distribution $P(x|\text{adult})$ or $P(x|$ "adult" is true), which are continuous. How do we obtain the denotation of "adult" (see Fig. 1)? With the denotation, e.g., the feature function or the membership function of class {Adult}, we can make new probability prediction or produce new likelihood function after $P(x)$ is changed. Can we obtain the feature function of set {Adult}? If the set {Adult} is fuzzy, can we obtain the membership function of the fuzzy set {Adult}?

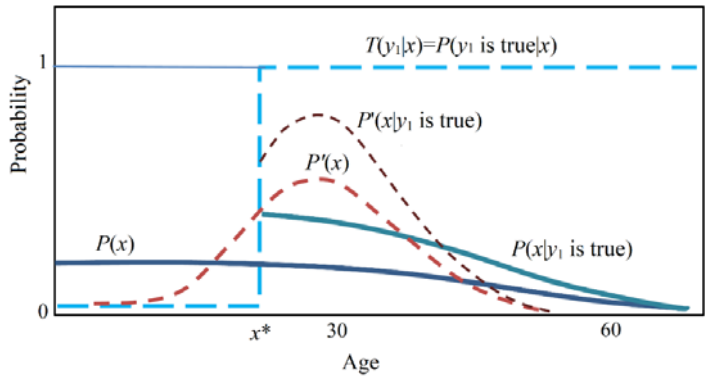

Fig. 1. Solving the denotation of "Adult" and the posterior distribution $P'(x|y_1 \text{ is true})$.

Further, if we only know a not big enough sample with unsmooth or discontinuous distribution $P(x, y)$, can we construct a smooth membership function with parameters as we do for a likelihood function, and train it with the sampling distribution? This is a label learning issue.

Furthermore, given a group of membership functions and a changed $P(x)$, how do we classify the instance space with maximum likelihood criterion or maximum mutual information criterion? This is a multi-label classification issue.

In this paper, to resolve above problems with membership functions, we first propose the third kind of Bayes' theorem and prove that the membership function obtained from

the new Bayes' formula is equivalent to that obtained from the random set statistics. Two resolve the above two issues, we use the semantic information method [11-13].

In next section, we introduce new mathematical methods. In Section 3, we simply introduce the applications of the new methods to multi-label classification, maximum mutual information classifications for unseen instances, and mixture models. Section 4 provides discussions. The last section is the summary.

2 Mathematical Methods

2.1 Distinguishing Statistical Probability and Logical Probability

Definition 1 Let U denote an instance set, and X denote a discrete random variable taking a value x from U. That means $X \in U=\{x_1, x_2, \ldots\}$. Let V denote the set of selectable labels, including some atomic and compound labels and let $Y \in V=\{y_1, y_2, \ldots\}$.

Definition 2 A label y_j is also a predicate $y_j(X)=$ "$X \in A_j$." For each y_j, U has a subset of A_j, every instance of 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 Logical Probability (LP) of y_j. For simplicity, let $P(y_j)= P(Y=y_j)$ and $T(y_j)=T(A_j)= P(X \in A_j)$.

We call $P(X \in A_j)$ the logical probability because according to Tarski's theory of truth [14], $P(X \in A_j)=P($"$X \in A_j$" is true$)=P(y_j$ is true). Hence the conditional LP of y_j for given X is the feature function of A_j and the truth function of y_j. We denote it with $T(A_j|X)$. Hence there is

$$T(A_j) = \sum_i P(x_i) T(A_j \mid x_i) \tag{2.1}$$

According to Davidson's truth-conditional semantics [15], $T(A_j|X)$ 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 $T(A_j)$ is also the fuzzy event probability defined by Zadeh [16]. For fuzzy sets, we use θ_j to replace A_j. Then $T(\theta_j|X)$ becomes the membership function of θ_j. There is

$$m_{\theta_j}(X) = T(\theta_j \mid X) = T(y_j \mid X) \tag{2.2}$$

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

2.2 The Three Kinds of Bayes' Theorems

There are three kinds of Bayes' theorem, which are used by Bayes [17], Shannon [18], and the author respectively.

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

$$T(B|A)=T(A|B)T(B)/T(A),\ T(A)= T(A|B)T(B)+ T(A|B^c)T(B^c) \tag{2.3}$$

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

Bayes' Theorem II (used by Shannon):

$$P(x_i \mid y_j) = P(y_j \mid x_i)P(x_i) / P(y_j),\ P(y_j) = \sum_i P(x_i)P(y_j \mid x_i) \tag{2.4}$$

There is also an asymmetrical formula for $P(y_j|x_i)$. Note there are two random variables and two statistical probabilities.

Bayes' Theorem III: Assume that $P(X)=P(X=\text{any in } U)$ and $T(\theta_j)=P(X \in \theta_j)$. Then

$$P(X \mid \theta_j) = T(\theta_j \mid X)P(X) / T(\theta_j),\ T(\theta_j) = \sum_i P(x_i)T(\theta_j \mid x_i) \tag{2.5}$$

$$T(\theta_j \mid X)=P(X \mid \theta_j)T(\theta_j) / P(X),\ T(\theta_j) = 1 / \max(P(X \mid \theta_j) / P(X)) \tag{2.6}$$

The two formulas are asymmetrical because there is a statistical probability and a logical probability. $T(\theta_j)$ in (2.5) may be call longitudinally normalizing constant.

The Proof of Bayes' Theorem III: Assume the joint probability $P(X,\ \theta_j)= P(X=\text{any}, X \in \theta_j)$, then $P(X|\theta_j)T(\theta_j) = P(X=\text{any}, X \in \theta_j)= T(\theta_j|X)P(X)$. Hence there is

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

Since $P(X|A_j)$ is horizontally normalized, $T(\theta_j)=\sum_i P(x_i)\ T(\theta_j|x_i)$. Since $T(\theta_j|X)$ is longitudinally normalized and has the maximum 1, 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.**

Equation (2.5) can be directly written in

$$T(\theta_j \mid X)=[P(X \mid \theta_j) / P(X)] / \max[P(X \mid \theta_j) / P(X)] \tag{2.7}$$

By this formula, we can obtain the denotation of "Adult" and the posterior distribution $P'(x|y_1$ is true) as shown in Fig. 1 where the set is not fuzzy.

2.3 Relationships between Likelihood functions, Membership Functions, and Sampling Distributions

In Shannon's information theory [18], $P(X)$ is called the source and $P(Y)$ is called the destination, the transition probability matrix $P(Y|X)$ is called the channel. A Shannon's channel consists of a group of transition probability functions: $P(y_j|X),\ j=1, 2, \ldots, n$.

$P(y_j|X)$ has two important properties: 1) It can be used for Bayes' prediction to get $P(X|y_j)$; after $P(X)$ becomes $P'(X)$, $P(y_j|X)$ still works; 2) $P(y_j|X)$ by a constant k can make the same probability prediction because

$$\frac{P'(X)kP(y_j \mid X)}{\sum_i P'(x_i)kP(y_j \mid x_i)} = \frac{P'(X)P(y_j \mid X)}{\sum_i P'(x_i)P(y_j \mid x_i)} = P'(X \mid y_j) \tag{2.8}$$

Similarly, a semantic channel consists of a group of membership functions: $T(\theta|X)$: $T(\theta_j|X)$, j=1, 2, …, n. According to (2.8), if $T(\theta_j|X) \propto P(y_j|X)$, there is $P(X|\theta_j)=P(X|y_j)$. Hence the optimized membership function is

$$T^*(\theta_j|X)=P(y_j|X)/\max(P(y_j|X)) \tag{2.9}$$

The relationships between membership functions, likelihood functions, and several probability distributions are:

$$\begin{aligned} T^*(\theta_j|X) &= [P^*(X|\theta_j)/P(X|)]/\max[P^*(X|\theta_j)/P(X|)] \\ &= [P(X|y_j)/P(X|)]/\max[P(X|y_j)/P(X|)]=P(y_j|X)/\max(P(y_j|X)) \end{aligned} \tag{2.10}$$

2.4 The Consistency of the Bayes' Theorem III and Random Set Statistics

We can prove that the membership function derived from (2.10) is the same as that from the random set statistics [2].

Assume that a Shannon channel $P(Y|X)$ is obtained from a big sample **D** where whose size is $N \to \infty$; X is equiprobable; there are N/m examples $(x_i; y_j)$ for every x_i with different y_j. We pick out all examples with y_j. Assume x' is the most instance in these examples. We denote it by x_j^*, whose number is N_j^*. Existing multi-instance multi-label learning method [19] reminds us that we can merge these examples with y_j into N_j^* multi-instance examples $S_k=(x_{k1}, x_{k2},\ldots; y_j)$, k=1, 2,…,N_j^*, every one of which contains x_j^*. Then we can treat S_k as a set-value taken by the random set. Let its feature function be denoted by $F_k(X)$.

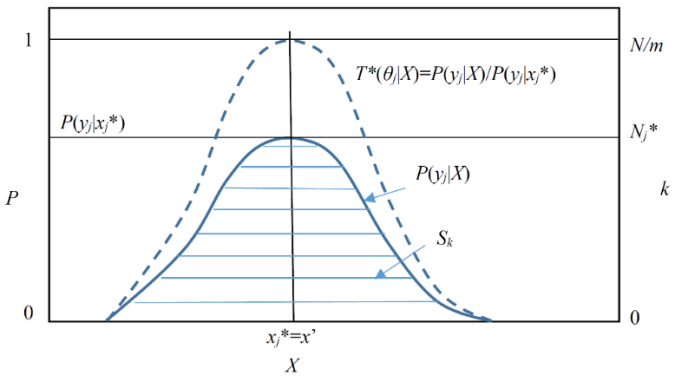

Fig. 2. The Bayes' Theorem III is compatible with the random sets falling shadow theory.

According to Wang's random set falling shadow theory [2], the membership function of θ_j is

$$m_{\theta_j}(X)=\frac{1}{N_j*}\sum_{k=1}^{N_j*}F_k(X) \tag{2.11}$$

According to classical statistics, the transition probability function of y_j is

$$P(y_j \mid X)=\frac{1}{(N/m)}\sum_{k=1}^{N_j*}F_k(X) \tag{2.12}$$

Comparing the above two formulas, we have

$$\begin{aligned} m_{\theta_j}(X) &= P(y_j \mid X)/[N_j*/(N/m)] \\ &= P(y_j \mid X)/\max[P(y_j \mid X)]=T*(\theta_j \mid X) \end{aligned} \tag{2.13}$$

If X is not equiprobable, we can randomly remove some examples to get an equiprobable sample. Its $P(Y|X)$ is the same, and hence the conclusion is the same.

2.5 Optimizing Membership Functions with Unsmooth or Discontinuous Sampling Distributions

If sampling distributions are unsmooth or discontinuous, but we wish that membership functions are smooth, then we can use the semantic information method to optimize the membership function.

The (amount of) semantic information conveyed by y_j about x_i is defined with log-normalized-likelihood [9]:

$$I(x_i;\theta_j)=\log\frac{P(x_i \mid \theta_j)}{P(x_i)}=\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{2.14}$$

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

$$I(x_i;\theta_j)=\log[1/T(\theta_j)]-(X-x_j)^2/(2d^2) \tag{2.15}$$

The $\log[1/T(\theta_j)]$ is Bar-Hillel-Carnap's semantic information measure [20]. Eq. (2.15) tells us that the larger the deviation is, the less information there is; the less the logical probability is, the more information there is; and, a wrong estimation may convey negative information. These conclusions accord with Popper's thought [21].

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

$$I(X;\theta_j)=\sum_i P(x_i \mid y_j)\log\frac{P(x_i|\theta_j)}{P(x_i)}=\sum_i P(x_i \mid y_j)\log\frac{T(\theta_j \mid x_i)}{T(\theta_j)} \tag{2.16}$$

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

$$T^*(\theta_j \mid X) = \arg\max_{T(\theta_j|X)} I(X;\theta_j) = \arg\max_{T(\theta_j|X)} \sum_i P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} \quad (2.17)$$

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 Kullback-Leibler information. When $P(y_j|X)$ is known and $P(X)$ is unknown, we may assume X is equiprobable to have

$$T^*(\theta_j \mid X) = \arg\max_{T(\theta_j|X)} I(X;\theta_j) = \arg\max_{T(\theta_j|X)} \sum_i \frac{P(y_j|x_i)}{\sum_k P(y_j|x_k)} \log \frac{T(\theta_j \mid x_i)}{\sum_k T(\theta_j|x_k)} \quad (2.18)$$

To average $I(x_i; \theta_j)$ in（2.15）for different X and Y, we have

$$\begin{aligned} I(X;\theta) &= H(\theta) - H(\theta \mid X) \\ &= -\sum_j P(y_j) \log T(\theta_j) - \sum_j \sum_i P(x_i, y_j)(x_i - x_j)^2 / \left(2d_j^{\,2}\right) \end{aligned} \quad (2.19)$$

It is easy to find that the Maximum Semantic Information (MSI) criterion is a special Regularized Least Squares (RLS) criterion. $H(\theta|X)$ is the mean squared error, and $H(\theta)$ is the negative regularization term.

3 Applications to Machine Learning

3.1 Multi-Label Learning and Classification

There have been many valuable studies in Multi-label learning and classification [22-24]. In popular methods, the learning and the classification are made by the same agent. However, from the viewpoint of semantic communication, the sender's classification and the receiver's logical classification are different. The receiver learns from a sample to obtain labels' denotations, e. g., membership functions, whereas the sender needs, for a given instance, to select a label with the most information. The sender partitions the instance space whereas the receiver does not.

Section 2.5 has discussed how to obtain optimized membership functions, which makes multi-label learning much easier because the learning is naturally converted into several single label learning. We may improve Binary Relevance method [24] to optimize the membership function of a label with both positive and negative examples by

$$\begin{aligned} T^*(\theta_j \mid X) &= \arg\max_{T(\theta_j|X)} [I(X;\theta_j) + I(X;\theta_j^c)] \\ &= \arg\max_{T(\theta_j|X)} \sum_i [P(x_i \mid y_j) \log \frac{T(\theta_j \mid x_i)}{T(\theta_j)} + P(x_i \mid y_j') \log \frac{1\text{-}T(\theta_j \mid x_i)}{1\text{-}T(\theta_j)}] \end{aligned} \quad (3.1)$$

where $T^*(\theta_j|x_i)$ is only affected by $P(y_j|X)$ and $P(y_{j'}|X)$. For a given label, this method divides all instances into three kinds: the positive, the negative, and the unclear. $T^*(\theta_j|x_i)$ is not affected by unclear instances. However, popular One-vs-Rest or Binary Relevance method [24, 25] divides all instances into two kinds: the positive and the negative, for every label, and hence it needs a lot of time to prepare samples.

This binary logical learning allows the second part of Eq. (3.1) to be 0. Any big enough sample with distribution $P(X, Y)$ may be used for the membership function.

Now we discuss multi-label classifications with maximum semantic information criterion. For a visible instance X, the label sender selects y_j by the classifier

$$y_j^*=h(X)=\arg\max_{y_j}\log I(\theta_j;x_i)=\arg\max_{y_j}\log\frac{T(\theta_j\mid X)}{T(\theta_j)} \tag{3.2}$$

This classifier produces a noiseless Shannon channel. Using $T(\theta_j)$ can overcome the class-imbalance problem [22]. If $T(\theta_j|X)\in\{0,1\}$, the above semantic information measure becomes Bar-Hillel and Carnap's information measure [20]; the classifier becomes

$$y_j^*=h(X)=\arg\max_{y_j \text{ with } T(A_j|X)=1}\log[1/T(A_j)]=\arg\min_{y_j \text{ with } T(A_j|X)=1}T(A_j) \tag{3.3}$$

It means that we should select a label with the least logical probability and hence with the richest connotation. The above classifier encourages us to select a compound label such as y_1 and y_2 and y_3' (' means negation). Unlike canonical Binary Relevance method [25], it does not add label "Adult" or "Non-youth" to an example with label "Old person". See [14] for details about the new method for multi-label classifications.

3.2 Maximum Mutual Information Classifications for Unseen Instances

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 $\mathbf{D}$, we can obtain $P(X, Z)$. If $\mathbf{D}$ is not big enough, we may use the likelihood method to obtain $P(X, Z)$ with parameters. The aim is to solve the optimal partition of C. The problem is that Shannon's channel is not fixed and also needs optimization. Hence, we need semi-supersized learning method. We may use the Channels' Matching (CM) iteration algorithm [12,13].

Let C_j be a subset of C and $y_j=f(Z|Z\in C_j)$. Hence $S=\{C_1, C_2, \ldots\}$ is a partition of C. Our aim is, for given $P(X, Z)$ from D, to find optimized S, which is

$$S^*=\arg\max_S I(X;\theta|S)=\arg\max_S\sum_j\sum_i P(C_j)P(x_i\mid C_j)\log\frac{T(\theta_j\mid x_i)}{T(\theta_j)} \tag{3.4}$$

First, we obtain the Shannon channel for given S:

$$P(y_j\mid X)=\sum_{z_k\in C_j}P(z_k\mid X),\ j=1,2,\ldots,n \tag{3.5}$$

From this Shannon's channel, we can obtain the semantic channel $T(\theta|X)$ in numbers or with parameters. For given Z, we have conditional semantic information

$$I(X_i;\theta_j|Z)=\sum_i P(X_i\mid Z)\log\frac{T(\theta_j\mid X_i)}{T(\theta_j)} \tag{3.6}$$

Then let the Shannon channel match the semantic channel by

$$y_j=f(Z)=\arg\max_{y_j} I(X;\theta_j\mid Z),\quad j=1,2,\ldots,n \tag{3.7}$$

Repeat (3.5)-(3.7) until S does not change. The convergent S is S^* we seek. Some iterative examples show that the above algorithm is fast and reliable. The convergence can be proved with the help of the $R(G)$ function [12].

3.3 Mixture Models

Assume a sampling distribution $P(X)$ is produced by two or several conditional probability functions $P^*(X|y_j)$ （j=1,2,…,n）, where $P^*(X|y_j)$ is some kind of function such as Gaussian distribution. We only know n, without knowing $P(Y)$. We need to find $P(Y)$ and model parameters θ so that the predicted distribution, denoted by $P_\theta(X)$, is as close to $P(X)$ as possible, e. g., the relative entropy or Kullback-Leibler divergence $H(P||P_\theta)=\sum_i P(x_i)\log[P(x_i)/P_\theta(x_i)]$ is close to 0.

The Expectation-Maximization (EM) algorithm and its improved versions [23, 24] are popular for solving mixture models. We can improve the EM algorithm or Maximization-Maximization algorithm [24] by the CM algorithm as follows:

Left-step a: Construct Shannon's channel by

$$\begin{aligned}&P(y_j\mid X)=P(y_j)P(X\mid\theta_j)/P_\theta(X)\\&P_\theta(X)=\sum_j P(y_j)(X\mid\theta_j)\end{aligned},\quad j=1,2,\ldots,n \tag{3.8}$$

This formula has been used in the E-step of the EM algorithm.

Left-step b Use the following equation to obtain a new $P(Y)$ repeatedly until the inner iteration converges:

$$\begin{aligned}P(y_j)&\Leftarrow\sum_i P(x_i)P(y_j\mid x_i)\\&=\sum_i P(x_i)\frac{P(x_i\mid\theta_j)}{\sum_k P(y_k)P(x_i\mid\theta_k)}P(y_j),\ j=1,2,\ldots,n\end{aligned} \tag{3.9}$$

If $H(P||P_\theta)$ is less than a small number, such as 0.001 bit, then end the iteration.

Right-step: Optimize the parameters in the likelihood function $P(X|\theta)$ on the right of the following log to maximize the semantic mutual information:

$$I(X;\theta)=\sum_i\sum_j P(x_i)\frac{P(x_i\mid\theta_j)}{P_\theta(x_i)}P(y_j)\log\frac{P(x_i\mid\theta_j)}{P(x_i)} \tag{3.10}$$

Then go to Left-step a.

Fortunately, to prove $H(P\|P_\theta)\to 0$, we derived an important formula [25]

$$\min H(P\,\|\,P_\theta)=\min_{P(Y),\,\theta}(I(X;Y)-I(X;\theta))=\min_{P(Y),\,\theta}(R(G)-G)) \tag{3.11}$$

where G is the semantic information and $R(G)$ is the minimum Shannon's mutual information for given G. In every step, $H(P\|\,P_\theta)$ is decreasing. In comparison with the EM algorithm, the CM algorithm has faster speed and clearer convergence reason [28].

The package with Excel files and Word files illustrating the CM algorithm for mixture models and maximum mutual information classifications can be obtained from http://survivor99.com/lcg/CM-iteration.zip. These Excel files also contain the data of iterative processes.

4 Discussions

4.1 The Significance to the Unification of Logic and Probability

It has been being an important issue to unify logic and probability [21, 28, 29]. Although logical probability has been discussed for a long time, people defined logical probability only with classical sets and hence truth functions can only be 0 or 1. In this way, it is hard to unify statistical probability and logical probability.

Zadeh proposed fuzzy sets and membership functions and explained a membership function as the truth function of a hypothesis [1]. This theory is an important advance because the truth function is between 0 and 1 and hence is closer to statistical probability. Peizhuang Wang [2] and others also made important advances in setting up relationship between statistics and fuzzy logic. However, there is still a gap between statistical probability and fuzzy logic because we cannot convert a sampling distribution such as $P(y_j|X)$ into a membership function reasonably. And, it is still unclear to derive membership functions from likelihood functions. It seems that the Bayes' Theorem III and the semantic information method can bridge this gap well.

4.2 From Bayesian Inference to Logical Bayesian Inference

In comparison with the likelihood function $P(X|\theta_j)$ and the Bayesian posterior $P(\theta|X)$ [30], the membership function $T(\theta_j|X)$ so obtained seems to be a better tool for machine learning because the main task of machine learning is classification. And, a fuzzy set indicates a fuzzy class and a membership function indicates the denotation of a label. The learning is just for the denotation of a label.

An important advantage of a group of membership functions, or a semantic channel, as a predictive model is that when the source $P(X)$ is changed, this model still works well and hence has good generalization performance [14].

The above methods for membership functions may be called Logical Bayesian Inference [31], an improved version of Bayesian inference.

5 Summary

This paper proposed and proved the third kind of Bayes' theorem including two asymmetrical formulas for transform between likelihood functions and the membership functions. Letting a semantic channel match a Shannon's channel, we can obtain a group of optimized membership functions from a sampling distribution. If the sampling distribution is unsmooth or discontinuous, we can obtain a group of optimized membership functions by the semantic information formula. The paper introduced the applications of new methods to multi-label classification, maximum mutual information classifications, and mixture models. It discussed the significance of the new methods for membership functions to the unification of logic and probability. As a new and general tool, Logical Bayesian Inference was proposed.

References

1. Zadeh, L. A.: Fuzzy Sets. Information and Control, 8, 338–53 (1965).
2. 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).
3. Dubois D, Moral S, Prade H. A Semantics for Possibility Theory Based on Likelihoods. Journal of Mathematical Analysis and Applications, 205, 359–380 (1997).
4. Thomas S F. Possibilistic uncertainty and statistical inference, ORSA/TIMS Meeting, Houston, Texas (1981).
5. Civanlar, M. R., Trussell, H. J.: Constructing membership functions using statistical data, Fuzzy Sets and Systems, 18, 1-13 (1986).
6. Pota, M., Esposito, M., De Pietro, G.: Transforming probability distributions into membership functions of fuzzy classes: A hypothesis test approach. Fuzzy Sets and Systems, 233, 52-73 (2013).
7. Lu., C.: B-fuzzy quasi-Boolean algebra and a generalize mutual entropy formula. Fuzzy Systems and Mathematics (in Chinese), 5, 76-80 (1991).
8. Lu, C.: A Generalized Information Theory (in Chinese). China Science and Technology University Press, Hefei (1993).
9. Lu, C.: Meanings of generalized entropy and generalized mutual information for coding, J. of China Institute of Communication(in Chinese), 15, 37-44 (1994).
10. Lu, C.: A generalization of Shannon's information theory. Int. J. of General Systems 28, 453-490 (1999).
11. 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).
12. 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).
13. Lu, C.: Semantic Channel and Shannon Channel Mutually Match for Multi-label Classification, in: Z. Shi et al. (Eds.) IFIP International Federation for Information Processing, ICIS 2018. pp 37-48. Springer Nature Switzerland AG: Switzerland (2018).

14. Tarski, A.: The semantic conception of truth: and the foundations of semantics. Philosophy and Phenomenological Research 4, 341–376 (1944).
15. Davidson, D.: Truth and meaning. Synthese 17, 304-323 (1967).
16. Zadeh, L. A.: Probability measures of Fuzzy events, Journal of Mathematical Analysis and Applications, 23, 421–427 (1986).
17. Bayes, T., Price, R.: An essay towards solving a problem in the doctrine of chance. Philosophical Transactions of the Royal Society of London 53, 370–418 (1763).
18. Shannon, C. E.: A mathematical theory of communication. Bell System Technical Journal 27, 379–429 and 623–656 (1948).
19. Zhou, Z. H., Zhang, M. L., Huang, S. J., Li, Y. F.: Multi-instance multi-label learning. Artificial Intelligence, 176, 2291–2320 (2012).
20. Bar-Hillel Y, Carnap R.: An outline of a theory of semantic information. Tech. Rep. No. 247, Research Lab. of Electronics, MIT (1952).
21. Popper, K.: Conjectures and Refutations. Repr. Routledge, London and New York (1963/2005).
22. Zhou, Z. H.: Machine Learning (in Chinese), Beijing: Tsinghua University Press, 2016.
23. Zhang, M. L., Zhou, Z. H.: A review on multi-label learning algorithm. IEEE Transactions on Knowledge and Data Engineering 26, 1819-1837(2014).
24. Zhang, M. L., Li, Y. K., Liu, X. Y., et al.: Binary Relevance for Multi-Label Learning: An Overview, Front. Comput. Sci. 12, 191–202(2018).
25. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1–38 (1977).
26. Neal, R., Hinton, G.: A view of the EM algorithm that justifies incremental, sparse, and other variants. in: Michael I. Jordan (ed.) Learning in Graphical Models, pp 355–368. MIT Press, Cambridge, MA (1999).
27. Lu, C.: Problems with the EM algorithm and the way out (in Chinese). http://survivor99.com/lcg/CM/Recent.html, last accessed 2018/4/10.
28. Jaynes, E. T. Probability Theory: The logic of Science, edited by Bretthorst, L., Cambridge University press, New York (2003).
29. Russell, S.: Unifying Logic and Probability, Communications of the ACM, 58, 88-97 (2015).
30. Lu, C.: From Bayesian inference to logical Bayesian inference: A new mathematical frame for semantic communication and machine learning. in： Z. Shi et al. (Eds.) IFIP International Federation for Information Processing, ICIS 2018, pp 11-23. Springer Nature Switzerland AG: Switzerland (2018).