Journal of Education and Educational Research
ISSN: 2957-9465 | Vol. 4, No. 2, 2023

Conditional Probability in Machine Learning

Xiaolei Jiang

School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China

Abstract: To help teaching of machine learning course, manipulation rules and application examples of conditional probabilities in machine learning are presented. The emphasis is to make a clear distinction between reasonable assumptions and logical deductions developed from assumptions and axioms. The formula for conditional probability of conditional probability is presented with examples in Bayesian coin tossing, Bayesian linear regression, and Gaussian processes for regression and classification. The signal + noise model is formulated in terms of a proposition and exemplified by linear-Gaussian models and linear dynamical systems.

Keywords: Conditional Probability; Bayesian Learning; Gaussian Processes for Regression and Classification; Linear-Gaussian Models; Linear Dynamical Systems.

1. Introduction

Machine learning is a core curriculum for undergraduate degree programs in artificial intelligence [1]. The probabilistic viewpoint and Bayesian methods play key roles in machine learning since many learning problems can be naturally formulated and solved in terms of probabilities [2]. For example, the basic problem of supervised learning can be described as finding the conditional probability of the target value given a training set. However, the applications of conditional probabilities in machine learning are ambiguous, vague, and difficult for students without a correct understanding [3]. We believe that only when a clear distinction between reasonable assumptions and logical deductions developed from assumptions has been made can the tool of conditional probabilities be fully mastered by students [4]. It is the goal of this article to expose the manipulation rules of conditional probabilities and give a plenty of examples about how these rules are applied in machine learning problems.

The rest of this article is organized as follows. After reviewing the definition of conditional probability, we give the formula for conditional probability of conditional probability in Section 2, by which the properties of probabilities can be easily rephrased in terms of conditional probabilities. In Section 3 examples in Bayesian learning are presented to illustrate how to use the manipulation rules given in Section 2. The signal + noise model is presented in Section 4, exemplified by linear-Gaussian models and linear dynamical systems. We draw our conclusion in Section 5.

2. Conditional Probability Review

It is well known that if A and C are events, then the conditional probability of A given C is

$$p(A|C) = \frac{P(AC)}{p(C)} \quad . \tag{1}$$

For convenience $p(A|C)$ is also denoted as $p_C(A)$. From Equation (1) it can be derived that the formula for conditional probability of conditional probability is

$$p_C(A|B) = p(A|BC) \quad . \tag{2}$$

Conditional probabilities obey the axioms of probabilities, so they are also probabilities, meaning that conditional probabilities share the same properties with probabilities. For example, from $p(AB) = p(B)\,p(A|B)$, we know $p_C(AB) = p_C(B)\,p_C(A|B)$, and by Equation (2) this can be rewritten as

$$p(AB|C) = p(B|C)\,p(A|BC). \tag{3}$$

For another example, from the fact that $p(AB) = p(A)\,p(B)$ if and only if $p(A|B) = p(A)$, we know that $p_C(AB) = p_C(A)\,p_C(B)$ if and only if $p_C(A|B) = p_C(A)$, and by Equation (2) this can be rephrased as

$$p(AB|C) = p(A|C)\,p(B|C) \text{ if and only if } p(A|BC) = p(A|C). \tag{4}$$

Conditional probabilities make sense not only for events but also for random variables, so we can write $p(x|C)$ and $p(A|x)$, and formulas involving conditional probabilities are still valid when events are replaced by random variables. For example, from Equation (3) it follows that

$$p(x, y|z) = p(x|z)\,p(y|x, z) \tag{5}$$

and from Equation (4) it follows that

$$p(x, y|z) = p(x|z)\,p(y|z) \text{ if and only if } p(x|y, z) = p(x|z). \tag{6}$$

It is worth noting that Equations (5) and (6) are used repeatedly in probabilistic machine learning, as can be seen in the following sections.

3. Bayesian Learning

Now that we have seen the rules for manipulating conditional probabilities, this section gives examples to show how these rules are employed in Bayesian learning.

3.1. Bayesian Coin Tossing

Let $\mathcal{D} = \{x_1, \ldots, x_m\}$ be the results of tossing a coin m times, where $x_i = 1$ (0) designates that we have a head (tail) in the ith tossing, and what we want is the outcome of the next tossing, which is denoted as x. Assume $p(x_1, \ldots, x_m|\mu) = \prod_{i=1}^{m} p(x_i|\mu) = \prod_{i=1}^{m} \mu^{x_i}(1-\mu)^{1-x_i}$, where μ is the parameter of the Bernoulli distribution. The predictive probability is given by

$$p(x = 1|\mathcal{D}) = \int_0^1 p(x = 1, \mu|\mathcal{D})\,\mathrm{d}\mu = \int_0^1 p(\mu|\mathcal{D})\,p(x = 1|\mathcal{D}, \mu)\,\mathrm{d}\mu$$
$$= \int_0^1 p(\mu|\mathcal{D})\,p(x = 1|\mu)\,\mathrm{d}\mu$$

Note that the second equality follows from Equation (5), and that the last equality holds because $p(x=1|\mathcal{D},\mu)=p(x=1|\mu)$, which is the consequence of

$$p(\mathcal{D},x=1|\mu)=p(\mathcal{D}|\mu)\,p(x=1|\mu) \tag{7}$$

as can be seen from Equation (6). So, the above derivation relies on the additional assumption of Equation (7), without which the predictive probability $p(x=1|\mathcal{D})$ cannot be simplified to its final form.

3.2. Bayesian Linear Regression

Suppose $t_1,\ldots,t_m$ are the target values corresponding to the input vectors $\mathbf{x}_1,\ldots,\mathbf{x}_m$, respectively, and our aim is to find the target value s corresponding to the input vector $\mathbf{z}$. The training set $\{t_1,\ldots,t_m\}$ can be denoted as $\mathbf{t}$. Let the prior for parameters be $p(\mathbf{w})=\mathcal{N}(\mathbf{w}|0,\alpha^{-1}\mathbf{I})$, and the likelihood be $p(\mathbf{t}|\mathbf{w})=\mathcal{N}(\mathbf{t}|\mathbf{\Phi w},\beta^{-1}\mathbf{I})$, where $\mathbf{w}$ is the parameter vector, $\mathbf{\Phi}$ is the design matrix. The predictive probability is given by

$$p(s|\mathbf{t})=\int p(s,\mathbf{w}|\mathbf{t})\,\mathrm{d}\mathbf{w}=\int p(\mathbf{w}|\mathbf{t})\,p(s|\mathbf{t},\mathbf{w})\,\mathrm{d}\mathbf{w}$$
$$=\int p(\mathbf{w}|\mathbf{t})\,p(s|\mathbf{w})\,\mathrm{d}\mathbf{w}$$

Note that the second equality holds as an instance of Equation (5), and the last equality follows from $p(s|\mathbf{t},\mathbf{w})=p(s|\mathbf{w})$. From Equation (6) it is clear that the hidden assumption $p(s,\mathbf{t}|\mathbf{w})=p(s|\mathbf{w})\,p(\mathbf{t}|\mathbf{w})$ has been added.

3.3. Gaussian Processes for Regression and Classification

First, we consider the regression problem introduced in Section 3.2. After choosing an appropriate kernel function $k(\cdot,\cdot)$, the model of Gaussian processes for regression is given by

$$p\left(\begin{bmatrix}s\\\mathbf{t}\end{bmatrix}\right)=\mathcal{N}\left(\begin{bmatrix}s\\\mathbf{t}\end{bmatrix}\middle|\begin{bmatrix}0\\\mathbf{0}\end{bmatrix},\begin{bmatrix}u+\beta^{-1} & \mathbf{k}^{\mathrm{T}}\\ \mathbf{k} & \mathbf{V}+\beta^{-1}\mathbf{I}\end{bmatrix}\right) \tag{8}$$

where $u=k(\mathbf{z},\mathbf{z})$, $\mathbf{k}=[k(\mathbf{x}_1,\mathbf{z})\ \cdots\ k(\mathbf{x}_m,\mathbf{z})]^{\mathrm{T}}$, $[\mathbf{V}]_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)$. To gain a better understanding of Equation (8), we introduce auxiliary random variables $b_1,\ldots,b_m,a$ corresponding to the input vectors $\mathbf{x}_1,\ldots,\mathbf{x}_m,\mathbf{z}$, respectively, and assume them jointly have a Gaussian distribution of the form

$$p\left(\begin{bmatrix}a\\\mathbf{b}\end{bmatrix}\right)=\mathcal{N}\left(\begin{bmatrix}a\\\mathbf{b}\end{bmatrix}\middle|\begin{bmatrix}0\\\mathbf{0}\end{bmatrix},\begin{bmatrix}u & \mathbf{k}^{\mathrm{T}}\\ \mathbf{k} & \mathbf{V}\end{bmatrix}\right) \tag{9}$$

where $\mathbf{b}=[b_1\ \cdots\ b_m]^{\mathrm{T}}$. The relationship between a, $\mathbf{b}$ and s, $\mathbf{t}$ is given by

$$\begin{cases}p(s|a)=\mathcal{N}(s|a,\beta^{-1})\\ p(\mathbf{t}|\mathbf{b})=\mathcal{N}(\mathbf{t}|\mathbf{b},\beta^{-1}\mathbf{I})\\ p(s,\mathbf{t},a,\mathbf{b})=p(a,\mathbf{b})\,p(s|a)\,p(\mathbf{t}|\mathbf{b}).\end{cases} \tag{10}$$

From Equations (9) and (10) it follows that Equation (8) holds, so the latter formulation is equivalent to the first one. But Equations (9) and (10) are much easier to appreciate. Here a and $\mathbf{b}$ behave as the latent variables, which encode the intuition that similar inputs result in similar outputs, and s and $\mathbf{t}$ are the observed variables, which are related to the latent variables via additive noise.

The benefits of Equations (9) and (10) over Equation (8) are even greater when we consider a classification problem [5], where the basic setting is similar to those of a regression problem, except that the target values $t_1,\ldots,t_m,s$ are either 1 or -1. We still suppose that Equations (9) and (10) hold, but with the first two of Equation (10) replaced by

$$\begin{cases}p(s|a)=\sigma(sa)\\ p(\mathbf{t}|\mathbf{b})=\prod_{i=1}^{m}\sigma(t_ib_i),\end{cases}$$

where $\sigma(\cdot)$ is the sigmoid function. It can be shown that Equations (9) and (10) implies the conditional independence property

$$p(\mathbf{t},s|a)=p(\mathbf{t}|a)\,p(s|a). \tag{11}$$

The predictive distribution is given by

$$p(s=1|\mathbf{t})=\int p(a|\mathbf{t})\,p(s=1|\mathbf{t},a)\,\mathrm{d}a$$
$$=\int p(a|\mathbf{t})\,p(s=1|a)\,\mathrm{d}a$$

Note that the last equality follows by applying Equation (6) to Equation (11).

4. The Signal + Noise Model

Suppose $\mathbf{y}=g(\mathbf{x})+\mathbf{n}$, where $\mathbf{x}$ denotes the original signal, $\mathbf{n}$ denotes the noise, and $\mathbf{y}$ is the observed signal, all of which are random vectors. So before measured as $\mathbf{y}$, the original signal $\mathbf{x}$ is first transformed through g, and then contaminated by $\mathbf{n}$. If we further assume that $\mathbf{n}$ obeys the Gaussian distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$, then many students cannot resist the temptation to draw the conclusion that $p(\mathbf{y}|\mathbf{x})=\mathcal{N}(\mathbf{y}|g(\mathbf{x}),\mathbf{I})$, because $\mathbf{y}$ is just $\mathbf{n}$ plus a constant $g(\mathbf{x})$ when $\mathbf{x}$ is given. In fact, this is not the whole truth, as described in the following proposition.

Proposition 1 If $\mathbf{y}=g(\mathbf{x})+\mathbf{n}$, where $g(\cdot)$ is differentiable, then $p(\mathbf{y}|\mathbf{x})=f(\mathbf{y}-g(\mathbf{x}))$ if and only if $p(\mathbf{n})=f(\mathbf{n})$, and $\mathbf{x}$ and $\mathbf{n}$ are independent.

Proof $\Rightarrow$ Let $\tilde{\mathbf{x}}=\mathbf{x}$, then the transformations between $(\tilde{\mathbf{x}},\mathbf{n})$ and $(\mathbf{x},\mathbf{y})$ are

$$\begin{cases}\tilde{\mathbf{x}}=\mathbf{x}\\ \mathbf{n}=\mathbf{y}-g(\mathbf{x})\end{cases}\quad\text{and}\quad\begin{cases}\mathbf{x}=\tilde{\mathbf{x}}\\ \mathbf{y}=g(\tilde{\mathbf{x}})+\mathbf{n}.\end{cases}$$

From

$$p(\mathbf{x},\mathbf{y})=p(\mathbf{x})\,p(\mathbf{y}|\mathbf{x})=p(\mathbf{x})f(\mathbf{y}-g(\mathbf{x}))=p(\tilde{\mathbf{x}})f(\mathbf{n})$$

it follows that

$$p(\tilde{\mathbf{x}},\mathbf{n})=p(\mathbf{x},\mathbf{y})\left|\frac{\partial(\mathbf{x},\mathbf{y})}{\partial(\tilde{\mathbf{x}},\mathbf{n})}\right|=p(\tilde{\mathbf{x}})f(\mathbf{n}).$$

Hence $p(\mathbf{n})=f(\mathbf{n})$, and $\tilde{\mathbf{x}}$ and $\mathbf{n}$ are independent. The independence between $\mathbf{x}$ and $\mathbf{n}$ follows immediately since $\tilde{\mathbf{x}}=\mathbf{x}$.

$\Leftarrow$ Let $\tilde{\mathbf{x}}=\mathbf{x}$, then the transformations between $(\tilde{\mathbf{x}},\mathbf{y})$ and $(\mathbf{x},\mathbf{n})$ are

$$\begin{cases}\tilde{\mathbf{x}}=\mathbf{x}\\ \mathbf{y}=g(\mathbf{x})+\mathbf{n}\end{cases}\quad\text{and}\quad\begin{cases}\mathbf{x}=\tilde{\mathbf{x}}\\ \mathbf{n}=\mathbf{y}-g(\tilde{\mathbf{x}}).\end{cases}$$

Since

$$p(\tilde{\mathbf{x}},\mathbf{y})=p(\mathbf{x},\mathbf{n})\left|\frac{\partial(\mathbf{x},\mathbf{n})}{\partial(\tilde{\mathbf{x}},\mathbf{y})}\right|=p(\mathbf{x})f(\mathbf{n})=p(\tilde{\mathbf{x}})f(\mathbf{y}-g(\mathbf{x})),$$

and $\tilde{\mathbf{x}}=\mathbf{x}$, we have $p(\mathbf{y}|\mathbf{x})=\frac{p(\tilde{\mathbf{x}},\mathbf{y})}{p(\tilde{\mathbf{x}})}=f(\mathbf{y}-g(\mathbf{x}))$.

4.1. Linear-Gaussian Models

Consider a Bayesian network over n vertices $x_1,\cdots,x_n$, in which all conditional probability densities are Gaussian. Specifically, the conditional probability density of x_i given its parent vertices pa_i is

$$p(x_i|\mathrm{pa}_i)=\mathcal{N}\left(x_i\middle|\sum_{j\in\mathrm{pa}_i}w_{ij}x_j+b_i,v_i\right),$$

where w_{ij}, b_i, and v_i are parameters. Introducing $\varepsilon_i=x_i-\left(\sum_{j\in\mathrm{pa}_i}w_{ij}x_j+b_i\right)$, then Proposition 1 tells us that $\{x_j\}_{j\in\mathrm{pa}_i}$ and ε_i are independent and $p(\varepsilon_i)=\mathcal{N}(\varepsilon_i|0,v_i)$. So, we can derive formulas for $\mathbb{E}[x_i]$ and $\mathrm{cov}[x_i,x_j]$ with no other assumptions being made [3].

4.2. Linear Dynamical Systems

Consider a linear dynamical system in which the transition and emission distributions are $p(\mathbf{z}_i|\mathbf{z}_{i-1}) = \mathcal{N}(\mathbf{z}_i|\mathbf{A}\mathbf{z}_{i-1}, \mathbf{\Gamma})$ and $p(\mathbf{x}_i|\mathbf{z}_i) = \mathcal{N}(\mathbf{x}_i|\mathbf{C}\mathbf{z}_i, \mathbf{\Sigma})$, respectively, where $\mathbf{z}_i$ is the latent variable and $\mathbf{x}_i$ is the observed variable. Letting $\mathbf{w}_i = \mathbf{z}_i - \mathbf{A}\mathbf{z}_{i-1}$ and $\mathbf{v}_i = \mathbf{x}_i - \mathbf{C}\mathbf{z}_i$, then from Proposition 1 it follows that $p(\mathbf{w}_i) = \mathcal{N}(\mathbf{w}_i|0, \mathbf{\Gamma})$, $p(\mathbf{v}_i) = \mathcal{N}(\mathbf{v}_i|0, \mathbf{\Sigma})$, $\mathbf{w}_i$ and $\mathbf{z}_{i-1}$ are independent, $\mathbf{v}_i$ and $\mathbf{z}_i$ are independent. So, the formulation of transition and emission distributions implies that noises are Gaussian [3].

5. Conclusion

The conditional probabilities in machine learning are explored through Bayesian learning and the signal + noise model, exemplified by problems in Bayesian coin tossing, Bayesian linear regression, Gaussian processes for regression and classification, linear-Gaussian models, and linear dynamical systems. Emphasis is placed on the separation of assumptions and conclusions so that the tool of conditional probabilities employed in machine learning is fitted into a logical framework. Once students clarify what are assumptions and what are logical conclusions deduced from axioms and assumptions, mystery associated with the use of conditional probabilities in machine learning will disappear, and they can make their own way in solving machine learning problems via the tool of conditional probabilities.

Acknowledgments

The author thanks his colleges and students for their kindness and support.

References

[1] Wenqian Sun and Xing Gao (2018). The construction of undergraduate machine learning course in the artificial intelligence era. International Conference on Computer Science & Education, p.62-65.

[2] Kevin Patrick Murphy (2022). Probabilistic machine learning: an introduction. MIT Press.

[3] Christopher M. Bishop (2006). Pattern recognition and machine learning. Springer.

[4] Athanasios Papoulis and S. Unnikrishna Pillai (2002). Probability, random variables, and stochastic processes (fourth edition). McGraw-Hill.

[5] Carl Edward Rasmussen and Christopher K. I. Williams (2006). Gaussian processes for machine learning. MIT Press.