Bayesian Machine Learning via Category Theory (Culbertson and Sturtz, [2013a])

Overview.

The Category $\mathcal{P}$

The core construction in this paper is the category $\mathcal{P}$ of conditional probabilities. The objects in $\mathcal{P}$ are countably generated measurable spaces (X, Σ_X) and each arrow between (X, Σ_X) and (Y, Σ_Y) is a "regular conditional probability". We can think of regular conditional probabilities as functions that map elements of X to probability measures over (Y, Σ_Y), but there are some additional measurability constraints as well. In $\mathcal{P}$ we compute the composition of two arrows $T \colon (X, \Sigma_X) \to (Y, \Sigma_Y)$ and $U \colon (Y, \Sigma_Y) \to (Z, \Sigma_Z)$ by marginalizing over Y. We define the marginalization as:

$$(U \circ T)(C|x) = \int_{y \in Y} U(C|y) dT_x \tag{1}$$

We can think of the marginalization over Y as being the expectation of the measure U with respect to the distribution on Y defined by T. $\mathcal{P}$ is equivalent to the Kleisli category of the Giry Monad, but the authors don't use this fact until the "Final Remarks" section.

Example: The easiest way to reason about $\mathcal{P}$ is to examine a discrete example. Consider the discrete objects $\mathbf{2}, \mathbf{3}$ representing the discrete measurable spaces with two and three elements respectively. Each arrow from $\mathbf{2}$ to $\mathbf{3}$ represents a different way that observing an element from $\mathbf{2}$ defines a probability distribution over $\mathbf{3}$. For example, some arrow A might represent the case where the elements of $\mathbf{2}$ reflect whether the roll of a die was odd/even and the elements of $\mathbf{3}$ indicate whether the roll was in $\{1\}$, $\{2, 3, 4, 5\}$, or $\{6\}$. In this case, A is a function that maps the element in $\mathbf{2}$ corresponding to "odd" to the following measure:

$$\{1\} \to 0, \{2, 3, 4, 5\} \to 0.67, \{6\} \to 0.33 \tag{2}$$

In order to represent joint distributions in $\mathcal{P}$, the author defines the following: given two arrows P, Q (probability measures) in $\mathcal{P}$ that are each conditioned on the same space, we can represent the joint distribution J of P and Q as a product morphism of P and Q. The following image shows an example of this for the case where P and Q are "absolute" probability measures (arrows from 1):

$$\begin{tikzcd}
 & 1 \arrow[dl, "P"'] \arrow[d, dashed, "J"] \arrow[dr, "Q"] & \\
X & X \times Y \arrow[l, "\delta_{\pi_X}"] \arrow[r, "\delta_{\pi_Y}"'] & Y
\end{tikzcd}$$

In the above figure δ_{π_X} and δ_{π_Y} are the dirac measures of π_X and π_Y, the set projection maps of $X \times Y$ defined as $\pi_X(x, y) = x$ and $\pi_Y(x, y) = y$. These measures serve as the projection maps for the product $X \times Y$ in $\mathcal{P}$.

Now since two probability distributions may have many joint distributions, J is not unique and $X \times Y$ is actually a weak product in $\mathcal{P}$. Therefore, in order to uniquely define J we also need information about the relationship between P and Q. One way to do this is to express Q as the composition of P and some conditional distribution $h \colon X \to Y$.

$$\begin{tikzcd}
 & 1 \arrow[dl, "P_X"'] \arrow[d, dashed, "J_h"] & \\
X \arrow[rr, bend right, "h"'] & X \times Y \arrow[l, dashed, "\delta_{\pi_X}"'] \arrow[r, dashed, "\delta_{\pi_Y}"] & Y
\end{tikzcd}$$

Then we can define:

$$\forall A \in \Sigma_X, \forall B \in \Sigma_Y \qquad J_h(A \times B) = \int_A h_B dP \tag{3}$$

In non-categorical terms, this is equivalent to the representation of $P(A \cap B) = P(B|A)P(A)$. Of course, we know that $P(A \cap B) = P(A|B)P(B)$ as well, so we can derive a $k\colon Y \longrightarrow X$ such that:

$$\forall A \in \Sigma_X, \forall B \in \Sigma_Y \qquad \int_A h_B dP = J(A \times B) = \int_B k_A dQ \tag{4}$$

The Bayesian Paradigm in $\mathcal{P}$

We can use the properties of products and this equation to derive joint distributions from conditionals and vice versa. The authors' construction of Bayesian probability uses this mechanism to define inference maps in terms of prior and sampling distributions, as well as to derive posterior distributions from data and inference maps.

$$\begin{tikzcd}
& 1 \arrow[dl, "P_H"'] \arrow[dl, dashed, bend left, "\mathcal{I} \circ \mu"] \arrow[dr, dashed, "\mu"] & \\
H \arrow[rr, "\mathcal{S}"] & & D \arrow[ll, dotted, bend left, "\mathcal{I}"]
\end{tikzcd}$$

The Bayesian paradigm is based on two measurable spaces: D, which represents data observations and is typically a Euclidian space, and H, which is a parameterization of the model output, or decision. For example, in a model where we are estimating the relative probabilities of two classes, H might be the discrete measurable space with two elements.

The prior probability over H is the absolute probability P_H, and the map S is the sampling distribution over D. We can use P_H and S to define the joint distribution $J\colon 1 \longrightarrow H \times D$ as $J(A \times B) = \int_A S_B dP_H$. If we define the absolute probability $P_D = S \circ P_H$ then we can also use the product equation to define the inference map $I\colon D \longrightarrow H$:

$$\forall A \in \Sigma_H, \forall B \in \Sigma_D \qquad \int_B I_A dP_D = J(A \times B) = \int_A S_B dP_H \tag{5}$$

In order to model the process of updating the prior P_H to a posterior, we first introduce an arrow to represent data collection. $\mu\colon 1 \longrightarrow D$ is a point mass probability measure on D. The arrow $I \circ \mu\colon 1 \longrightarrow H$ is then the posterior distribution, which represents an improved "estimate" of the distribution over H in light of the measurement μ.

Stochastic Processes

Let us define the function space Y^X to be the set of all measurable functions from X to Y and define Σ_{Y^X} to be the σ-algebra over Y^X induced by the set of all point evaluation maps $\{ev_x\}_{x \in X}$ (where we define $ev_x(f) = f(x)$). Then (Y^X, Σ_{Y^X}) is an object in $\mathcal{P}$ and the authors define a stochastic process to be a $\mathcal{P}$-arrow $1 \longrightarrow Y^X$ (an absolute probability measure over a function space). Note that this is closely related to the non-categorical definition of stochastic processes: an X-indexed collection of distributions over Y can also be viewed as a distribution over functions that map the index set X to Y.

For most of the paper the authors limit their analysis to the case where Y is $\mathbb{R}$ and X is $\mathbb{R}^n$ and use a Gaussian process as their canonical example of a stochastic process. They define a Gaussian process to be a probability measure on Y^X such that for any finite subset $X_0 \subset X$ the induced probability measure over Y^{X_0} is a multivariate Gaussian distribution. We can write this distribution as $\mathcal{N}(m|_{X_0}, k|_{X_0})$, where the mean function $m|_{X_0} \in Y^{X_0}$ and the covariance function $k|_{X_0}\colon X_0 \times X_0 \longrightarrow Y$.

Stochastic Processes for Estimation

In Bayesian Machine Learning, we aim to construct a probability distribution on Y based on a data point $\mathbf{x}$. We can use our system to express this. For a given data point $\mathbf{x}$, we define the hypothesis space H to be Y^X and the data space D to be Y.

$$\begin{tikzcd} & 1 \arrow[dl, "P \sim \mathcal{GP}(m,k)"'] \arrow[dr, "Pev_{\mathbf{x}}^{-1} \sim \mathcal{N}(m(\mathbf{x}),k(\mathbf{x},\mathbf{x}))"] & \\ Y^X \arrow[rr, shift left, "\mathcal{S}^{\mathbf{x}} = \delta_{ev_{\mathbf{x}}}"] & & Y \arrow[ll, shift left, dashed, "\mathcal{I}^{\mathbf{x}}"] \end{tikzcd}$$

Then the sampling distribution S_x is the evaluation map of Y^X at $\mathbf{x}$ and the composition $Pev_{\mathbf{x}}^{-1}$ represents the probability distribution over Y given the point $\mathbf{x}$. We can build on this to incorporate model training (updating P with new data) by defining the joint distribution $J\colon 1 \to Y \times Y^X$ and using the product equation to derive the inference map:

$$\forall A \in \Sigma_{Y^X}, \forall B \in \Sigma_Y \qquad \int_{f \in A} S^{\mathbf{x}}(B|f)dP = J(B \times A) = \int_{y \in B} I^{\mathbf{x}}(A|y)d(Pev_{\mathbf{x}}^{-1}) \tag{6}$$

Then given an observation $(\mathbf{x}^*, y)$, we can derive the posterior distribution on Y^X as the stochastic process $I^{\mathbf{x}^*} \circ \delta_y = I^{\mathbf{x}^*}(\cdot|y)$. The adjusted probability distribution on Y given the point $\mathbf{x}$ is then $S^{\mathbf{x}} \circ I^{\mathbf{x}^*} \circ \delta_y$. In the paper the authors use this equation along with the definition of a Gaussian process to derive the following updates. Given a measurement $(\mathbf{x}, y)$, and a Gaussian process $\mathcal{GP}(m,k)$ we update the mean function m and covariance function k as follows:

$$m^1(\mathbf{z}) = m(\mathbf{z}) + \frac{k(\mathbf{z},\mathbf{x})}{k(\mathbf{x},\mathbf{x})}(y - m(\mathbf{x})) \tag{7}$$

$$k^1(\mathbf{w},\mathbf{z}) = k(\mathbf{w},\mathbf{z}) - \frac{k(\mathbf{w},\mathbf{x})k(\mathbf{x},\mathbf{z})}{k(\mathbf{x},\mathbf{x})} \tag{8}$$

Near the end of the paper the authors introduce a more general definition of a stochastic process as a point in the category $\mathcal{P}^X$ where X is any category. That is, a stochastic process is a natural transformation between the functor that maps all $x \in_{ob} X$ to 1 (the terminal object in $\mathcal{P}^X$) and some functor F in P^X. We can see that if X is a discrete category with no non-identity arrows then the set of functors F in P^X is isomorphic to the set of functions Y^X for $Y \in_{ob} \mathcal{P}$ and this definition reduces to the previous definition of a stochastic process as a $\mathcal{P}$-arrow $1 \to Y^X$.

We can use this more general definition to characterize a Markov Process. Consider the case where X is a total linear ordering $(T, \leq)$ and $\mathcal{F}$ is some functor in $\mathcal{P}^{(T, \leq)}$. Then given a sequence of ordered points $\{t_0, t_1, \ldots\}$ in T, their image under $\mathcal{F}$ is the following sequence where each $\mathcal{F}_{t_i, t_{i+1}} = \mathcal{F}(\leq)$ is a $\mathcal{P}$-arrow:

$$\mathcal{F}(t_1) \xrightarrow{\mathcal{F}_{t_1,t_2}} \mathcal{F}(t_2) \xrightarrow{\mathcal{F}_{t_2,t_3}} \mathcal{F}(t_3) \xrightarrow{\mathcal{F}_{t_3,t_4}} \ldots \tag{9}$$

By functorality and the definition of composition in $\mathcal{P}$ we can write:

$$\mathcal{F}_{t_i,t_{i+2}}(B|x) = \int_{u \in \mathcal{F}(t_{i+1})} \mathcal{F}_{t_{i+1},t_{i+2}}(B|u)d\mathcal{F}_{t_i,t_{i+1}}(\cdot|x) \tag{10}$$

Outside of category theory this is known as the Chapman-Kolomogorov relation, which defines the "memoryless" property of a Markov Process.

Comments.

Overall, I found this paper well constructed and original. However, I have a few comments.

First, although the paper does construct a formulation of inference maps in terms of their relationship to the other components of the Bayesian model, the authors do not describe any general procedure for computing inference maps. Their derivation of inference maps for Gaussian Processes leans heavily on the specific characteristics of Gaussian Processes rather than the categorical machinery that they develop throughout the rest of the paper. I believe that a full categorical formulation of Bayesian Machine Learning will require a formulation for deriving inference maps as well.

Next, this paper was very heavy on new and confusing notation, which made it hard to follow at points. The authors also overrode notation liberally. For example, the authors use Γ to describe both the graph of a function and the graph of a probability measure.

In addition, the authors only briefly mention the relationship of $\mathcal{P}$ to the Giry Monad. This is surprising given that the authors claim in their previous paper [2] that algebras over the Giry Monad could be the key to defining decision making on top of a probabilistic framework (which is the heart of machine learning).

Finally, the generalized definition of a stochastic process as a point in the category $\mathcal{P}^X$ seems like a promising way to define families of stochastic processes. I was surprised that the authors only mentioned how we could use this definition to characterize Markov Processes and Gaussian Processes, but did not dig deeper.

REFERENCES

[1] Culbertson, J. and Sturtz, K. (2013a). Bayesian machine learning via category theory. *arXiv preprint arXiv:1312.1445*.

[2] Culbertson, J. and Sturtz, K. (2013b). A categorical foundation for bayesian probability. *Applied Categorical Structures*, 22(4):647–662.