Parameter Learning and Prediction with Bayesian Belief Networks

Georg Heimel and Viktor Atalla

February 20, 2018

Abstract

Some thoughts on the matter.

Chapter 1

One Random Variable

1.1 Parameter or parameter distribution?

Suppose we have a single, discrete random variable X with probability mass function $P(X)$. The trivial belief network illustrating this variable is shown in Fig. 1.1. Strictly speaking, this graph illustrates the situation where $P(X)$ is

Figure 1.1: Belief network of a single, discrete random variable X with probability mass function $P(X)$.

given in free, unparameterized form, one number for each discrete value x that X can take. If we want to visually represent a *parameterized* probability mass function with parameter θ, then we have to introduce another node for it into the graph, as shown in Fig. 1.2. However, because the nodes in belief networks

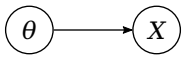

Figure 1.2: A single random variable X, whose probability mass function is parameterized by θ, which is itself a random variate.

represent probability mass functions (or densities), not numbers, this graph represents a joint probability distribution $P(X, \theta)$ of *two* random variables, not just one. In fact, what we have drawn designates θ as a (usually continuous) random variate as well. Furthermore, the graph encodes a particular factorization of the joint probability distribution.

$$P(X, \theta) = P(X|\theta)P(\theta) \tag{1.1}$$

To recover the (now *marginal*) probability mass function $P(X)$, we would

have to integrate out θ.

$$P(X) = \int \mathrm{d}\theta \, P(X, \theta) = \int \mathrm{d}\theta \, P(X|\theta) P(\theta) \tag{1.2}$$

In oder to do that, however, we would need to know and specify the probability density $P(\theta)$. If, in contrast, we just wanted to express the concept that $P(X)$ is a probability mass function parameterized by a *known*, well-defined value θ^*, then we have to add one more element to the graph, as shown in Fig. 1.3. Because θ^* is a constant and no probability distribution can be assigned to it,

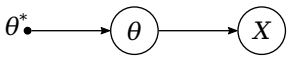

Figure 1.3: A single random variable X, whose probability mass function is parameterized by θ, which is set to the number θ^*.

above graph implies the following factorization of the overall joint probability.

$$P(X, \theta) = P(X|\theta) P(\theta|\theta^*) \tag{1.3}$$

In order to *set* the value of θ to θ^*, we chose a *delta distribution* centered on θ^* for the last term on the right-hand side.

$$P(\theta|\theta^*) = \delta(\theta - \theta^*) \tag{1.4}$$

Substituting 1.4 into Eq. 1.3 and integrating out θ yields:

$$P(X) = \int \mathrm{d}\theta \, P(X|\theta) \, \delta(\theta - \theta^*) = P(X|\theta^*) \tag{1.5}$$

Trying to visually represent the result of this integration, *i.e.*, the very right-hand side term, leads to the graph in Fig. 1.4. It literally spells out the parameterized probability mass function $P(X|\theta^*)$.

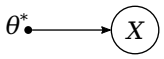

Figure 1.4: A single random variable X, whose probability mass function is parameterized by the *a priori* known and fixed number θ^*.

Returning to Eq. 1.2, we could also keep θ as a random variate but chose a parameterized distribution for it as well. Calling the parameters, again fixed and known numbers, α and β, the graph in Fig. 1.5 then encodes the factorization:

$$P(X, \theta) = P(X|\theta) P(\theta|\alpha, \beta) \tag{1.6}$$

If we wanted, for example, to learn the distribution of θ from the distribution of X with our *prior* belief encoded in $P(\theta|\alpha, \beta)$, we would factorize the left-hand side of Eq. 1.6 as $P(\theta|X)P(X)$ and divide by $P(X)$ to arrive at:

$$P(\theta|X) = \frac{P(X|\theta) P(\theta|\alpha, \beta)}{P(X)} \tag{1.7}$$

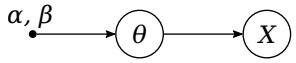

Figure 1.5: A single random variable X, whose probability mass function is parameterized by θ, itself given by a distribution parameterized by α and β.

1.2 Learning the parameter

Clearly, if we want to predict the outcome of a future *realization* of X, we need some idea of what the number θ^* or the probability density $P(\theta)$ should be. To learn either, we will first have to experiment and observe a number N of realizations x of X and the learned parameter (or its distribution) will depend on the specific outcome of our experiment. Trying to preserve a strict one-to-one correspondence between formulas and graphs, the repeated observations are modeled as a random process, where there are, in fact, N separate random variates $X_{\mathrm{i=1\ldots N}}$, all with the same probability mass function and all parameterized by the very same θ, which we take to be non-parametrically distributed to keep things simple for now.

We start with a trivial factorization of the joint probability distribution of all these random variables,

$$P(\theta|X_1, X_2, \ldots, X_{\mathrm{N}})P(X_1, X_2, \ldots, X_{\mathrm{N}}) = P(\theta, X_1, X_2, \ldots, X_{\mathrm{N}}) \tag{1.8}$$

and divide by the second term on the left-hand side.

$$P(\theta|X_1, X_2, \ldots, X_{\mathrm{N}}) = \frac{P(\theta, X_1, X_2, \ldots, X_{\mathrm{N}})}{P(X_1, X_2, \ldots, X_{\mathrm{N}})} \tag{1.9}$$

If each X_i is independent of all others *conditioned* on θ, that is, if the observations are *i.i.d*, we draw the graph for what we are about to see in our repeated experiment as shown in Fig. 1.6; the fact that circles representing the (conditional) probability mass functions of the X_{i} are shaded indicates that these are *observable* variables, while θ is a *hidden* variable. Using the factorization implied by the graph for the joint probability distribution in the numerator of the

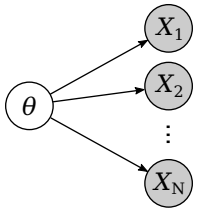

Figure 1.6: Multiple, independent instances of a single, observable random variable X, whose probability mass functions are all parameterized by the same θ, which is itself a random variate, but with a non-parametric distribution.

right-hand side of Eq. 1.9, we obtain:

$$P(\theta|X_1, X_2, \ldots, X_\mathrm{N}) = \frac{P(\theta)\prod_{i=1}^{N} P(X_\mathrm{i}|\theta)}{P(X_1, X_2, \ldots, X_\mathrm{N})} \tag{1.10}$$

Now we start our experiment and observe actual values for X, that is, first we observe that X_1 takes on the value x_1, $X_2 = \mathsf{x}_2$ and so on until we see that $X_\mathrm{N} = \mathsf{x}_\mathrm{N}$. Replacing all variables by their observed values in Eq. 1.10, leads us to:

$$P(\theta|\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_\mathrm{N}) = \frac{P(\theta)\prod_{i=1}^{N} P(\mathsf{x}_\mathrm{i}|\theta)}{P(\mathsf{x}_1, \mathsf{x}_2, \ldots, \mathsf{x}_\mathrm{N})} \tag{1.11}$$

The set of observed realizations $\{\mathsf{x}_\mathrm{i}\}$ is commonly referred to as *the data*, denoted by $\mathcal{D}$. Using this notation, we finally obtain:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \tag{1.12}$$

Now, given a suitable expression for $P(\mathsf{x}_i|\theta)$, the *likelihood* of θ given the observation x_i, and a choice for the *prior* probability distribution $P(\theta)$, we can learn the *posterior* probability distribution of θ. In general, however, computing the denominator in Eq. 1.12, the *evidence* for the data, is quite cumbersome. Therefore, one often resorts to taking the value θ^*, where the product $P(\theta|\mathcal{D})P(\theta)$ has a maximum, as a *point estimate* for θ. Specifically, we take the *maximum likelihood estimate* (MLE) if we chose a flat prior $P(\theta) = const.$

$$\theta_\mathrm{MLE} = \mathrm{argmax}_\theta\ P(\mathcal{D}|\theta) \tag{1.13}$$

and the *maximum posterior* (MAP) estimate if we don't.

$$\theta_\mathrm{MAP} = \mathrm{argmax}_\theta\ P(\mathcal{D}|\theta)P(\theta) \tag{1.14}$$

Mathematically, this again corresponds to *collapsing* the posterior to a delta distribution $\delta(\theta - \theta^*)$ with its center θ^* either θ_MLE or θ_MAP. Depending on whether we do or do not chose to replace the posterior distribution with a point estimate, we end up with the graphs shown in Figs. 1.4 and 1.2, respectively, with the difference that we now have *learned* the parameter θ (or its probability density), and not just set it.

Note that, if we wanted to express our prior belief of $P(\theta)$ with a parameterized probability density, we would extend all graphs in the present Section like we demonstrated in Fig. 1.5 and replace $P(\theta)$ by $P(\theta|\alpha, \beta)$ in all formulas accordingly (*cf.* Eq. 1.7).

1.3 Predicting on new data

1.3.1 The short answer

We can now proceed to predict the probability mass function for the outcome of a new realization of our univariate random process. To distinguish this event from

past observations, we introduce yet another random variate, X^* this time. Now, if we chose a (learned) point estimate θ^* for the parameter of the probability mass function (*e.g.*, MLE or MAP), we have to look no further than Fig. 1.4 and Eq. 1.5.

$$P(X^*) = P(X^*|\theta^*) \tag{1.15}$$

If, however, we decided to keep the entire posterior distribution of θ, then we are in the situation depicted in Fig. 1.2 and, consequently, have to employ Eq. 1.2.

$$P(X^*) = \int \mathrm{d}\theta \, P(X^*|\theta)P(\theta) \tag{1.16}$$

1.3.2 The long answer

To explicitly express that $P(\theta)$ is indeed our *posterior* belief about the distribution of θ, and that we are indeed predicting the probability mass function for a new observation *after* having seen the data, we explicitly introduce X^* into the graph in Fig. 1.6 to produce the one shown in Fig. 1.7. Collecting all the X_i

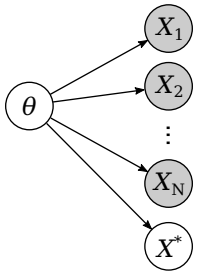

Figure 1.7: Same as Figure 1.6 with X^* representing a new, unobserved instance of X.

as $\mathcal{D}$, the joint probability of all three parameters X^*, θ, and $\mathcal{D}$, can then be factorized in two ways which are, of course, equal to each other. Specifically, we chose:

$$P(X^*, \theta|\mathcal{D})P(\mathcal{D}) = P(X^*, \mathcal{D}|\theta)P(\theta) \tag{1.17}$$

Because all observations of X, past or future, are independent of each other conditioned on θ, X^* is also independent from $\mathcal{D}$. We can, therefore, rewrite the right-hand side,

$$P(X^*, \theta|\mathcal{D})P(\mathcal{D}) = P(X^*|\theta)P(\mathcal{D}|\theta)P(\theta) \tag{1.18}$$

join its last two terms,

$$P(X^*, \theta|\mathcal{D})P(\mathcal{D}) = P(X^*|\theta)P(\mathcal{D}, \theta) \tag{1.19}$$

take them apart again the other way around,

$$P(X^*, \theta|\mathcal{D})P(\mathcal{D}) = P(X^*|\theta)P(\theta|\mathcal{D})P(\mathcal{D}) \tag{1.20}$$

and divide by $P(\mathcal{D})$.

$$P(X^*, \theta|\mathcal{D}) = P(X^*|\theta)P(\theta|\mathcal{D}) \tag{1.21}$$

Now integrating over θ brings us to the formal expression for the *posterior predictive*.

$$P(X^*|\mathcal{D}) = \int \mathrm{d}\theta \, P(X^*|\theta)P(\theta|\mathcal{D}) \tag{1.22}$$

Chapter 2

Two Random Variables

2.1 Repeat after me ...

Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be three non-intersecting sets of nodes in the directed graph G, that is $\mathcal{X} \cap \mathcal{Y} = \mathcal{X} \cap \mathcal{Z} = \mathcal{Y} \cap \mathcal{Z} = \varnothing$. We define $\mathcal{X}$ and $\mathcal{Y}$ to be d-separated by $\mathcal{Z}$ in G iff all paths between every $x \in \mathcal{X}$ and every $y \in \mathcal{Y}$ are *blocked*. A path U is said to be blocked iff there is a node w on U such that either

- w is a collider and neither w nor any of its descendants is in $\mathcal{Z}$, or
- w is not a collider on U and w is in $\mathcal{Z}$.

If, in particular, G represents a *belief network* and, thus, the nodes represent random variates, then $\mathcal{X}$ and $\mathcal{Y}$ are independent conditional on $\mathcal{Z}$ in all probability distributions G can represent iff $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$.

2.2 Parameter separation

Moving on, we now suppose we have two random variables, X and Y, parameterized by θ_{x} and θ_{y}, respectively. To make things more interesting, we will further suppose that they are not independent of each other but that, on the contrary, the value that X takes impacts the distribution of values observed for Y. Because our distributions are parameterized, affecting the probability mass function of Y means affecting the parameter θ_{y} that defines it. The graph encoding the situation just described, shown in Fig. 2.1, suggest the following factorization of the joint distribution of all involved random variates.

$$P(Y, X, \theta_{\mathrm{y}}, \theta_{\mathrm{x}}) = P(Y|\theta_{\mathrm{y}}, X)P(\theta_{\mathrm{y}}|X)P(X|\theta_{\mathrm{x}})P(\theta_{\mathrm{x}}) \tag{2.1}$$

In an attempt to find a second, complementary factorization that would allow us to find the analogue to Eq. 1.12 for two parameters, we start by conditioning on X.

$$P(Y, \theta_{\mathrm{y}}, \theta_{\mathrm{x}}|X)P(X) = rhs \tag{2.2}$$

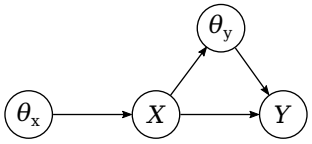

Figure 2.1: Two correlated random variables, X and Y, with their probability mass functions parameterized by θ_x and θ_y, which are themselves random variables, but with non-parametric distributions.

Using the recipe from Section 2.1, we then realize that both Y and θ_y are independent of θ_x conditioned on X (because X is on any path between them, is *not* a collider, and *is* in the conditioning set).

$$P(Y, \theta_y|X)P(\theta_x|X)P(X) = rhs \tag{2.3}$$

Next, we combine the lone $P(X)$ with the first factor,

$$P(Y, X, \theta_y)P(\theta_x|X) = rhs \tag{2.4}$$

and tear that factor apart in a different way.

$$P(\theta_y|Y, X)P(Y, X)P(\theta_x|X) = rhs \tag{2.5}$$

Writing out again the right-hand side of Eq. 2.1 and dividing by $P(Y, X) = P(Y|X)P(X)$ then yields:

$$P(\theta_y|Y, X)P(\theta_x|X) = \frac{P(Y|\theta_y, X)P(\theta_y|X)P(X|\theta_x)P(\theta_x)}{P(Y|X)P(X)} \tag{2.6}$$

Comparing this to Eq. 1.12, we see that we got what we asked for, sort of. But how are we going to learn the parameters θ_x and θ_y at the same time? The answer is, we aren't because Eq. 2.6 is actually the product of two equations, one for each parameter, and we need to separate it to learn them one by one.

First, we combine the last to terms in the numerator on the right-hand side of Eq. 2.6 to $P(\theta_x, X)$ and take that apart the other way around.

$$P(\theta_y|Y, X)P(\theta_x|X) = \frac{P(Y|\theta_y, X)P(\theta_y|X)P(\theta_x|X)P(X)}{P(Y|X)P(X)} \tag{2.7}$$

Then we realize that the newly created factors cancel with existing ones only to yield,

$$P(\theta_y|Y, X) = \frac{P(Y|\theta_y, X)P(\theta_y|X)}{P(Y|X)} \tag{2.8}$$

which we recognize as the "first half" of Eq. 2.6. We will use this expression to learn the X-dependent parameter θ_y.

To isolate an equivalent expression for θ_x, we start again with Eq. 2.6 and multiply with $P(X)$,

$$P(\theta_y|Y, X)P(X)P(\theta_x|X) = \frac{P(Y|\theta_y, X)P(\theta_y|X)P(X)P(X|\theta_x)P(\theta_x)}{P(Y|X)P(X)} \tag{2.9}$$

combine the first three terms in the numerator on the right-hand side,

$$P(\theta_{\mathrm{y}}|Y,X)P(X)P(\theta_{\mathrm{x}}|X) = \frac{P(Y,X,\theta_{\mathrm{y}})P(X|\theta_{\mathrm{x}})P(\theta_{\mathrm{x}})}{P(Y|X)P(X)} \tag{2.10}$$

multiply by $P(Y|X)$,

$$P(\theta_{\mathrm{y}}|Y,X)P(Y|X)P(X)P(\theta_{\mathrm{x}}|X) = \frac{P(Y,X,\theta_{\mathrm{y}})P(X|\theta_{\mathrm{x}})P(\theta_{\mathrm{x}})}{P(X)} \tag{2.11}$$

and combine the first three terms on the left-hand side,

$$P(Y,X,\theta_{\mathrm{y}})P(\theta_{\mathrm{x}}|X) = \frac{P(Y,X,\theta_{\mathrm{y}})P(X|\theta_{\mathrm{x}})P(\theta_{\mathrm{x}})}{P(X)} \tag{2.12}$$

which cancel with their right-hand side equivalent to give us,

$$P(\theta_{\mathrm{x}}|X) = \frac{P(X|\theta_{\mathrm{x}})P(\theta_{\mathrm{x}})}{P(X)} \tag{2.13}$$

which we recognize as the "second half" of Eq. 2.6.

2.3 Learning the parameters

2.3.1 The short answer

Looking at Eq. 2.2, we realize that we can proceed in exactly the same manner as for the univariate case discussed in Section 1.2 to learn the parameter θ_{x}. We denote the obtained posterior $P(\theta_{\mathrm{x}}|\mathcal{D}_{\mathrm{x}})$ because we do not, in fact, need the values observed for Y to learn θ_{x}, but only the set $\{\mathsf{x}_{\mathrm{i}}\} = \mathcal{D}_{\mathrm{x}}$ of realizations of X. To learn θ_{y}, we take another look at Eq. 2.8 and realize that everything is conditioned on X, by definition a *discrete* random variable that can take on the values $x_1, x_2, \ldots$ So, instead of just one equation, we are actually looking at as many as there are different values for X.

$$\begin{aligned} P(\theta_{\mathrm{y}}|Y,X=x_1) &= \frac{P(Y|\theta_{\mathrm{y}},X=x_1)P(\theta_{\mathrm{y}}|X=x_1)}{P(Y|X=x_1)} \\ P(\theta_{\mathrm{y}}|Y,X=x_2) &= \frac{P(Y|\theta_{\mathrm{y}},X=x_2)P(\theta_{\mathrm{y}}|X=x_2)}{P(Y|X=x_2)} \\ &\vdots \end{aligned} \tag{2.14}$$

This suggest that, having collected paired realizations of X and Y, we group by the possible values of X, specify a prior $P(\theta_{\mathrm{y}}|X)$ for each of them, and compute a separate posterior for θ_{y} from the Y realizations in each group, again following Section 1.2.

2.3.2 The long answer

Let's start by staring at Eq. 2.8 yet again.

$$P(\theta_{\mathrm{y}}|Y,X) = \frac{P(Y|\theta_{\mathrm{y}},X)P(\theta_{\mathrm{y}}|X)}{P(Y|X)}$$

It was derived by rigorously applying the rules for manipulating probability distributions but, to actually use it, we now must understand just what each term really means. Starting with the first of the two factors in the numerator, we acknowledge that it gives us the possibility to specify a different likelihood for each value x that X can take. While we might not actually want to do that, we still note that the product

$$P(Y|\theta_{\mathrm{y}},X) = \prod_k p_{\mathrm{k}}(Y|\theta_{\mathrm{y}})^{[X=x_{\mathrm{k}}]} \tag{2.15}$$

does the trick of filtering out the likelihood $p_{\mathrm{k}}(Y|\theta_{\mathrm{y}})$ by setting X to x_{k}. This is thanks to the *Iverson bracket* in the exponent, which evaluates to 1 if its contents are true and to 0 otherwise.

Likewise, the second term in the numerator intuitively expresses that a different prior $\pi(\theta_{\mathrm{y}})$ can be chosen for every value x that X can take. This is again expressed in the same way,

$$P(\theta_{\mathrm{y}}|X) = \prod_k \pi_{\mathrm{k}}(\theta_{\mathrm{y}})^{[X=x_{\mathrm{k}}]} \tag{2.16}$$

which yields for the entire numerator:

$$P(Y|\theta_{\mathrm{y}},X)P(\theta_{\mathrm{y}}|X) = \prod_k \Big\{ p_{\mathrm{k}}(Y|\theta_{\mathrm{y}})\pi_{\mathrm{k}}(\theta_{\mathrm{y}}) \Big\}^{[X=x_{\mathrm{k}}]} \tag{2.17}$$

Finally, the denominator is the integral of the numerator over θ_{y}.

$$\begin{aligned} P(Y|X) &= \int \mathrm{d}\theta_{\mathrm{y}} \prod_k \Big\{ p_{\mathrm{k}}(Y|\theta_{\mathrm{y}})\pi_{\mathrm{k}}(\theta_{\mathrm{y}}) \Big\}^{[X=x_{\mathrm{k}}]} \\ &= \prod_k \left\{ \int \mathrm{d}\theta_{\mathrm{y}}\, p_{\mathrm{k}}(Y|\theta_{\mathrm{y}})\pi_{\mathrm{k}}(\theta_{\mathrm{y}}) \right\}^{[X=x_{\mathrm{k}}]} \\ &\stackrel{!}{=} \prod_k p_{\mathrm{k}}(Y)^{[X=x_{\mathrm{k}}]} \end{aligned} \tag{2.18}$$

Putting it all together, then leads us to realize that the conditional posterior

$$\begin{aligned} P(\theta_{\mathrm{y}}|Y,X) &= \prod_k \left\{ \frac{p_{\mathrm{k}}(Y|\theta_{\mathrm{y}})\pi_{\mathrm{k}}(\theta_{\mathrm{y}})}{p_{\mathrm{k}}(Y)} \right\}^{[X=x_{\mathrm{k}}]} \\ &\stackrel{!}{=} \prod_k p_{\mathrm{k}}(\theta_{\mathrm{y}}|Y)^{[X=x_{\mathrm{k}}]} \end{aligned} \tag{2.19}$$

is actually a product of as many posteriors as there a values that X can take. Our goal, of course, is to learn every single one of them.

Extending the graph

Having grasped the multitude of posteriors we want to learn as well as their functional relationships, we now take a step back, and extend the graph in Fig. 2.1 as we did in Section 1.2 to include multiple instances of the two random variates that are to be observed. Note that the monstrosity in Fig. 2.2 seems

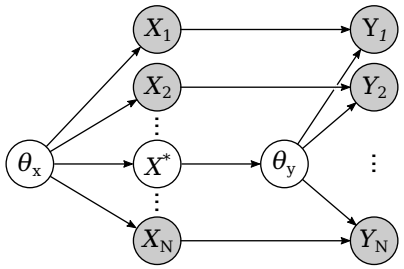

Figure 2.2: Multiple, independent instances of two correlated random variables, X and Y, with their probability mass functions parameterized by θ_x and θ_y, which are themselves random variables, but with non-parametric distributions.

to lie somewhere between Figs. 1.6 and 1.7. We will see soon enough that the extra, unobserved X is necessary to capture the conditional dependence of the posterior in θ_y.

This graph then encodes the following factorization of the joint probability of all involved variables (*cf.* Eq. 2.1):

$$lhs = \left\{\prod_{i=1}^{N} P(Y_i|\theta_y, X_i)\right\} P(\theta_y|X^*) \left\{\prod_{i=1}^{N} P(X_i|\theta_x)\right\} P(X^*|\theta_x)\, P(\theta_x) \tag{2.20}$$

To find the matching left-hand side, we first redo the exact same derivation as in Section 2.2, with the entire set of all $\{Y_i\}$ taking the place of the lone Y and, likewise, the entire set of all $\{X_i\}$ plus the X^* substituted for the lone X. This first leads us to the equivalent of Eq. 2.5.

$$P(\theta_y|\{Y_i\}, \{X_i\}, X^*)\, P(\theta_x|\{X_i\}, X^*)\, P(\{Y_i\}, \{X_i\}, X^*) = rhs \tag{2.21}$$

Also the next step is analogous to the preceding Section 2.2 in that we factorize the last term into $P(\{Y_i\}|\{X_i\}, X^*)\, P(\{X_i\}, X^*)$, which allows us again to separate into two parts.

$$P(\theta_y|\{Y_i\}, \{X_i\}, X^*) = \frac{P(\theta_y|X^*) \prod_{i=1}^{N} P(Y_i|\theta_y, X_i)}{P(\{Y_i\}|\{X_i\}, X^*)} \tag{2.22}$$

$$P(\theta_x|\{X_i\}, X^*) = \frac{P(X^*|\theta_x) P(\theta_x) \prod_{i=1}^{N} P(X_i|\theta_x)}{P(\{X_i\}, X^*)} \tag{2.23}$$

Posterior of θ_{x}

The appearance of the auxiliary quantity X in Eq. 2.23 is sitting somewhat oddly in the above expression for the posterior of θ_{x}. Why should one of the instances of X stick out from the rest? Fortunately, it is not hard to make it disappear. To do so, we multiply with the denominator of the right-hand side,

$$P(\theta_{\mathrm{x}}, \{X_{\mathrm{i}}\}, X^*) = P(X^*|\theta_{\mathrm{x}})\, P(\theta_{\mathrm{x}}) \prod_{i=1}^{N} P(X_{\mathrm{i}}|\theta_{\mathrm{x}}) \tag{2.24}$$

and pluck apart the left-hand side by conditioning on θ_{x}.

$$P(\{X_{\mathrm{i}}\}, X^*|\theta_{\mathrm{x}})\, P(\theta_{\mathrm{x}}) = rhs \tag{2.25}$$

Examining the graph in Fig. 2.2 in the light of Section 2.1, we realize that paths from X^* to any of the $\{X_{\mathrm{i}}\}$ either lead us through θ_{x}, which is not a collider but is in the conditioning set, or through one of the $\{Y_{\mathrm{i}}\}$, which are all colliders and are not in the conditioning set. Thus, all these paths are blocked and we can further write:

$$\begin{aligned} P(\{X_{\mathrm{i}}\}|\theta_{\mathrm{x}})\, P(X^*|\theta_{\mathrm{x}})\, P(\theta_{\mathrm{x}}) &= rhs \\ P(\theta_{\mathrm{x}}, \{X_{\mathrm{i}}\})\, P(X^*|\theta_{\mathrm{x}}) &= rhs \\ P(\theta_{\mathrm{x}}|\{X_{\mathrm{i}}\})\, P(\{X_{\mathrm{i}}\})\, P(X^*|\theta_{\mathrm{x}}) &= rhs \end{aligned} \tag{2.26}$$

Now that we see how the offensive term $P(X^*|\theta_{\mathrm{x}})$ cancels with its equivalent on the right-hand side, we can divide again by $P(\{X_{\mathrm{i}}\})$ and replace the variables by their realizations $\{\mathsf{x}_{\mathsf{i}}\}$ to arrive at:

$$P(\theta_{\mathrm{x}}|\{\mathsf{x}_{\mathsf{i}}\}) = \frac{P(\theta_{\mathrm{x}}) \prod_{i=1}^{N} P(\mathsf{x}_{\mathsf{i}}|\theta_{\mathrm{x}})}{P(\{\mathsf{x}_{\mathsf{i}}\})} \tag{2.27}$$

If we denote, as before, $\{\mathsf{x}_{\mathsf{i}}\} = \mathcal{D}_{\mathrm{x}}$, we finally end up with the familiar expression for the posterior of θ_{x}.

$$P(\theta_{\mathrm{x}}|\mathcal{D}_{\mathrm{x}}) = \frac{P(\mathcal{D}_{\mathrm{x}}|\theta_{\mathrm{x}}) P(\theta_{\mathrm{x}})}{P(\mathcal{D}_{\mathrm{x}})} \tag{2.28}$$

Posterior(s) of θ_{y}

To make sense of Eq. 2.22, we first plug Eqs. 2.15 and 2.16 into the numerator (num) of the right-hand side.

$$num = \prod_{j} \pi_{\mathrm{j}}(\theta_{\mathrm{y}})^{[X^*=x_{\mathrm{k}}]} \prod_{i=1}^{N} \left\{ \prod_{k} p_{\mathrm{k}}(Y_{\mathrm{i}}|\theta_{\mathrm{y}})^{[X_{\mathrm{i}}=x_{\mathrm{k}}]} \right\} \tag{2.29}$$

Rearranging the products, we get,

$$num = \prod_k \left\{ \pi_{\mathrm{k}}(\theta_{\mathrm{y}}) \prod_{i=1}^{N} p_{\mathrm{k}}(Y_{\mathrm{i}}|\theta_{\mathrm{y}})^{[X_{\mathrm{i}}=x_{\mathrm{k}}]} \right\}^{[X^*=x_{\mathrm{k}}]}$$
$$\stackrel{!}{=} \prod_k \Big\{ p_{\mathrm{k}}(\{Y_{\mathrm{i}}\}\,|\theta_{\mathrm{y}}, \{X_{\mathrm{i}}\}) \pi_{\mathrm{k}}(\theta_{\mathrm{y}}) \Big\}^{[X^*=x_{\mathrm{k}}]} \tag{2.30}$$

which leaves us with

$$den = \prod_k \left\{ \int \mathrm{d}\theta_{\mathrm{y}}\, \pi_{\mathrm{k}}(\theta_{\mathrm{y}}) \prod_{i=1}^{N} p_{\mathrm{k}}(Y_{\mathrm{i}}|\theta_{\mathrm{y}})^{[X_{\mathrm{i}}=x_{\mathrm{k}}]} \right\}^{[X^*=x_{\mathrm{k}}]}$$
$$\stackrel{!}{=} \prod_k \Big\{ p_{\mathrm{k}}(\{Y_{\mathrm{i}}\} \mid \{X_{\mathrm{i}}\}) \Big\}^{[X^*=x_{\mathrm{k}}]} \tag{2.31}$$

for the denominator (*den*). Putting it all together and substituting the actual realizations, $\{\mathsf{y}_{\mathrm{i}}\}$ and $\{\mathsf{x}_{\mathrm{i}}\}$, of Y and X, respectively, we then arrive at:

$$P(\theta_{\mathrm{y}}|\{\mathsf{y}_{\mathrm{i}}\}, \{\mathsf{x}_{\mathrm{i}}\}, X^*) = \prod_k \left\{ \frac{p_{\mathrm{k}}(\{\mathsf{y}_{\mathrm{i}}\}\,|\theta_{\mathrm{y}}, \{\mathsf{x}_{\mathrm{i}}\}) \pi_{\mathrm{k}}(\theta_{\mathrm{y}})}{p_{\mathrm{k}}(\{\mathsf{y}_{\mathrm{i}}\} \mid \{\mathsf{x}_{\mathrm{i}}\})} \right\}^{[X^*=x_{\mathrm{k}}]}$$
$$\stackrel{!}{=} \prod_k \Big\{ p_{\mathrm{k}}(\theta_{\mathrm{y}}|\,\{\mathsf{y}_{\mathrm{i}}\}\,, \{\mathsf{x}_{\mathrm{i}}\}) \Big\}^{[X^*=x_{\mathrm{k}}]} \tag{2.32}$$

Introducing again $\mathcal{D}_{\mathrm{y}} = \{\mathsf{y}_{\mathrm{i}}\}$, $\mathcal{D}_{\mathrm{x}} = \{\mathsf{x}_{\mathrm{i}}\}$, and $\mathcal{D} = \{\mathsf{y}_{\mathrm{i}}, \mathsf{x}_{\mathrm{i}}\}$, we could also rewrite this as,

$$P(\theta_{\mathrm{y}}|\mathcal{D}, X) = \prod_k \Big\{ p_{\mathrm{k}}(\theta_{\mathrm{y}}|\mathcal{D}) \Big\}^{[X^*=x_{\mathrm{k}}]} \tag{2.33}$$

where the relevant factors are given by:

$$p_{\mathrm{k}}(\theta_{\mathrm{y}}|\mathcal{D}) = \frac{p_{\mathrm{k}}(\mathcal{D}_{\mathrm{y}}|\theta_{\mathrm{y}}, \mathcal{D}_{\mathrm{x}}) \pi_{\mathrm{k}}(\theta_{\mathrm{y}})}{p_{\mathrm{k}}(\mathcal{D}_{\mathrm{y}}|\mathcal{D}_{\mathrm{x}})} \tag{2.34}$$

Proceed at your own risk from here!

2.4 Predicting on new data

To predict the probability mass function for a new observation X^*, we proceed in exactly the same way as for the single-variable case outlined in Section 1.3 and rewrite Eq. 1.22.

$$P(X^*|\mathcal{D}) = \int \mathrm{d}\theta_{\mathrm{x}}\, P(X^*|\theta_{\mathrm{x}}) P(\theta_{\mathrm{x}}|\mathcal{D}) \tag{2.35}$$

The same is true if we wanted to predict the probability mass function for a new observation Y^* given a certain value of X^*.

$$P(Y^*|X^*, \mathcal{D}) = \int \mathrm{d}\theta_\mathrm{y}\, P(Y^*|\theta_\mathrm{y}, X^*) P(\theta_\mathrm{y}|\mathcal{D}, X^*) \tag{2.36}$$

Again, Eq. 2.36 is actually many equations, one for every value x^* that X^* can take. Multiplying with Eq. 2.35,

$$P(Y^*, X^*|\mathcal{D}) = P(Y^*|X^*, \mathcal{D}) P(X^*|\mathcal{D}) \tag{2.37}$$

we get the joint posterior predictive,

$$P(Y^*, X^*|\mathcal{D}) = \int \mathrm{d}\theta_\mathrm{y} \mathrm{d}\theta_\mathrm{x}\, P(Y^*|\theta_\mathrm{y}, X^*) P(\theta_\mathrm{y}|\mathcal{D}, X^*) P(X^*|\theta_\mathrm{x}) P(\theta_\mathrm{x}|\mathcal{D}) \tag{2.38}$$

and that for a new observation X^* given a certain value of Y^*.

$$P(X^*|Y^*, \mathcal{D}) = \frac{P(Y^*, X^*|\mathcal{D})}{P(Y^*|\mathcal{D})} \tag{2.39}$$

The denominator on the right-hand side of this expression is obtained by summing the joint posterior predictive in Eq. 2.38 over all values $\{x_\mathrm{k}\}$ that X^* can possibly take.

$$P(Y^*|\mathcal{D}) = \sum_k P(Y^*, x_\mathrm{k}|\mathcal{D}) \tag{2.40}$$

We note in passing that, if the discrete random variable X is relabeled C and "... has the value x." is substituted with "... belongs to class c.", then Eq. 2.39 describes the recipe for predicting the *class probabilities* of new observation Y^*, and *classification* would mean assigning it to the class c^* with the highest probability.

2.5 Hyperparameters

Now that we understand the case of two dependent, parameterized random variables, we might want to include the possibility of specifying the (prior) distributions of these parameters with *hyperparameters*. Like in Fig. 1.5, adding two of them (why not α and β) for θ_x is straightforward. When we want to express our prior belief about θ_y, however, we should reserve the option to specify different parameters (why not μ and ν) for each value x that X can take. This means adding an arrow from X to μ and ν, as shown in Fig. 2.3.

Without going through every step of our derivation again, we list here only how introducing hyperparameters changes a few key formulas.

2.6 Bayesian hypothesis test

Suppose X is a binary random variable, representing if a patient has recieved a certain treatment or not, and Y is also binary random variable, indicating

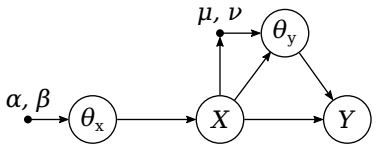

Figure 2.3: Two correlated random variables, X and Y, with their probability mass functions parameterized by θ_{x} and θ_{y}, which are themselves random variables defined through their respective hyperparamters (α, β) and (μ, ν).

whether patient's symptoms disappeared or not. Our goal is then to learn the distributions of the parameters, θ_{x} and θ_{y}, describing the probabilities of X and Y occurring, respectively.

As we are dealing with binary variables, we can chose between a *Beta-Bernoulli* and a *Beta-Binomial* model. As the names imply, the priors on θ_{x} and θ_{y} are *Beta* distributions in both models, each characterized by two parameters.

Chapter 3

Good Things Come in 3's

Example binary classification

Chapter 4

Continuous Variables

Density vs. mass function and the mystery of observation and hidden variables.