Notes on Bayesian Regression

Jens Behley

April 7, 2021

1 Notation

Some words on the notation:

- Vectors $\mathbf{x} \in \mathbb{R}^D$, $\mathbf{y} \in \mathbb{R}^{1\times D}$ and matrices $\mathbf{X} \in \mathbb{R}^{M\times N}$ are bold. Vectors are usually column vectors if not defined explicitly as row vectors, like $\mathbf{y} \in \mathbb{R}^{1\times D}$
- Sets are calligraphic letters: $\mathcal{X}, \ldots$, where the corresponding letter $x_i \in \mathcal{X}$ usually refers to elements from the set.
- Blackboard bold letters refer to the usual domains: $\mathbb{R}$ are real numbers, $\mathbb{N}$ are natural numbers including 0.
- We write $\mathbf{x}_{1:N}$ or $y_{1:N}$, when we refer to $\mathbf{x}_1, \ldots, \mathbf{x}_N$ or $y_1, \ldots, y_N$, respectively.
- We generally refer to $\mathbf{x} \in \mathbb{R}^D$ for the feature vector and y for the label, where $y \in \mathbb{R}$ for regression and $y \in \{1, \ldots, K\}$ for classification. We use $\theta \in \mathbb{R}^D$ for the parameters.
- For the normal distribution aka Gaussian, we use $\mathcal{N}(x|\mu, \sigma^2)$ to refer to a normal distribution with mean μ and variance σ^2.

2 Bayesian Regression

As the derivation of the posterior is a bit involved, here some notes on the derivation. Here, we try to make all steps explicit.

2.1 Derivation of the Posterior $P(\theta|y_{1:N}, \mathbf{x}_{1:N})$

Let's start with writing out the posterior:

$$P(\theta|y_{1:N}, \mathbf{x}_{1:N}) = \frac{P(y_{1:N}|\mathbf{x}_{1:N}, \theta)P(\theta)}{P(y_{1:N}|\mathbf{x}_{1:N})}, \tag{1}$$

where $\mathbf{x}_n \in \mathbb{R}^D$, $y_n \in \mathbb{R}$, and $\theta \in \mathbb{R}^K$.

Note that the marginal likelihood $P(y_{1:N}|\mathbf{x}_{1:N})$ can be written as:

$$\begin{aligned} P(y_{1:N}|\mathbf{x}_{1:N}) &= \int P(y_{1:N}, \theta|\mathbf{x}_{1:N})\ \mathrm{d}\theta \qquad (2)\\ &= \int P(y_{1:N}|\theta, \mathbf{x}_{1:N}) P(\theta|\mathbf{x}_{1:N})\ \mathrm{d}\theta \qquad (3)\\ &= \int P(y_{1:N}|\theta, \mathbf{x}_{1:N}) P(\theta)\ \mathrm{d}\theta, \qquad (4)\end{aligned}$$

where we exploited that $P(\theta, \mathbf{x}_{1:N}) = P(\theta)P(x_{1:N})$ due to the independence of θ and $\mathbf{x}_{1:N}$. Thus, the marginal likelihood is an integral of an product of the likelihood and the prior.

Therefore, we now concentrate on the numerator of Eq. 1. As before, let the likelihood $P(y_{1:N}|\mathbf{x}_{1:N}, \theta)$ and prior $P(\theta)$ given by normal distributions:

$$\begin{aligned} P(y_{1:N}|\mathbf{x}_{1:N}, \theta) &= \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\theta, \sigma^2\mathbf{Id}) \qquad (5)\\ P(\theta) &= \mathcal{N}(\theta|\mathbf{m}_0, \mathbf{S}_0) \qquad (6)\end{aligned}$$

Note that we have multivariate Gaussians and therefore the likelihood has a diagonal covariance $\sigma^2\mathbf{Id}$.

Plugging this into the numerator of Eq. 1, we get:

$$P(y_{1:N}|\mathbf{x}_{1:N}, \theta)P(\theta) = \mathcal{N}(\mathbf{y}|\mathbf{\Phi}\theta, \sigma^2\mathbf{Id})\mathcal{N}(\theta|\mathbf{m}_0, \mathbf{S}_0) \qquad (7)$$

First, we make a change of variables for the likelihood, where we switch roles of the variable $\mathbf{y}$ and the mean $\mathbf{\Phi}\theta$. We exploit the following theorem (see Prince, Chap. 5.7):

Theorem 1 *Let be $\mathcal{N}(\mathbf{x}|\mathbf{A}\mathbf{y} + \mathbf{b}, \Sigma)$ a Gaussian over $\mathbf{x}$ with a mean given by a linear transformation of another variable $\mathbf{y}$. Then we can re-express the Gaussian in terms of $\mathbf{y}$ as follows:*

$$\mathcal{N}(\mathbf{x}|\mathbf{A}\mathbf{y} + \mathbf{b}, \Sigma) = \eta^{-1}\mathcal{N}(\mathbf{y}|\mathbf{A}'\mathbf{x} + \mathbf{b}', \Sigma'), \qquad (8)$$

where

$$\begin{aligned} \Sigma' &= (\mathbf{A}^T\Sigma^{-1}\mathbf{A})^{-1} \qquad (9)\\ A' &= \Sigma'\mathbf{A}^T\Sigma^{-1} \qquad (10)\\ b' &= -\mathbf{A}'\mathbf{b} \qquad (11)\\ \eta &= |\mathbf{A}| \qquad (12)\end{aligned}$$

The derivation of the normalization constant can be found in the repository of Joe Dinius[1].

[1] https://github.com/jwdinius/prince-computer-vision/blob/master/Ch5/Prince-Chapter-5.ipynb.

Using Theorem 1, we get the following:

$$\mathcal{N}(\mathbf{y}|\mathbf{\Phi}\theta, \sigma^2\mathbf{Id}) = \eta_1^{-1}\mathcal{N}(\theta|\mathbf{A}'\mathbf{y}, \Sigma')$$

with

$$\Sigma' = (\mathbf{\Phi}^T(\sigma^2\mathbf{Id})^{-1}\mathbf{\Phi})^{-1} = (\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi})^{-1} \tag{13}$$

$$\mathbf{A}' = \Sigma'\Phi^T(\sigma^2\mathbf{Id})^{-1} = (\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\sigma^{-2}\Phi^T \tag{14}$$

We use η_1 as the normalization constant as this will later can be removed. Now, we can multiply the likelihood and prior using the following theorem (see Deisenroth et al., Chap. 6.5.2):

Theorem 2 *Assume we have two Gaussians over* $\mathbf{x} \in \mathbb{R}^D$*,* $\mathcal{N}(\mathbf{x}|\mathbf{a}, \mathbf{A})$ *and* $\mathcal{N}(\mathbf{x}|\mathbf{b}, \mathbf{B})$*. Then the product of the Gaussians is given by:*

$$\mathcal{N}\left(\mathbf{x}|\mathbf{a}, \mathbf{A}\right)\mathcal{N}(\mathbf{x}|\mathbf{b}, \mathbf{B}) = \eta^{-1}\mathcal{N}(\mathbf{x}|\mathbf{c}, \mathbf{C}), \tag{15}$$

where

$$\mathbf{C} = \left(\mathbf{A}^{-1} + \mathbf{B}^{-1}\right)^{-1} \tag{16}$$

$$\mathbf{c} = \mathbf{C}\left(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b}\right) \tag{17}$$

$$\eta = \mathcal{N}\left(\mathbf{a}|\mathbf{b}, \mathbf{A} + \mathbf{B}\right) = \mathcal{N}(\mathbf{b}|\mathbf{a}, \mathbf{A} + \mathbf{B}) \tag{18}$$

Replacing the likelihood with the changed variable Gaussian and applying the product theorem, Theorem 2, we get the following:

$$\begin{aligned}
&\mathcal{N}\left(\mathbf{y}|\mathbf{\Phi}\theta, \sigma^2\mathbf{Id}\right)\mathcal{N}\left(\theta|\mathbf{m}_0, \mathbf{S}_0\right) && (19)\\
&= \eta_1^{-1}\mathcal{N}\left(\theta|(\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\sigma^{-2}\Phi^T\mathbf{y}, (\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\right)\mathcal{N}(\theta|\mathbf{m}_0, \mathbf{S}_0) && (20)\\
&= \eta_1^{-1}\eta_2^{-1}\mathcal{N}\left(\theta|\mathbf{m}_N, \mathbf{S}_N\right) && (21)\\
& && (22)
\end{aligned}$$

with

$$\begin{aligned}
\mathbf{S}_N &= \left(\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{S}_0^{-1}\right)^{-1} && (23)\\
\mathbf{m}_N &= \mathbf{S}_N\left(\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi}(\sigma^{-2}\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\sigma^{-2}\mathbf{\Phi}^T\mathbf{y} + \mathbf{S}_0^{-1}\mathbf{m}_0\right) && (24)\\
&= \mathbf{S}_N\left(\sigma^{-2}\mathbf{\Phi}^T\mathbf{y} + \mathbf{S}_0^{-1}\mathbf{m}_0\right) && (25)
\end{aligned}$$

That means up to normalization constants η_1 and η_2, the product of the likelihood and prior is a Gaussian.

When we now put this into the marginal likelihood $P(y_{1:N}|\mathbf{x}_{1:N})$, we get the following:

$$
\begin{aligned}
P(y_{1:N}|\mathbf{x}_{1:N}) &= \int P(y_{1:N}|\theta, \mathbf{x}_{1:N}) P(\theta) \, \mathrm{d}\theta && (26)\\
&= \int \eta_1^{-1} \eta_2^{-1} \mathcal{N}\left(\theta|\mathbf{m}_N, \mathbf{S}_N\right) \, \mathrm{d}\theta && (27)\\
&= \eta_1^{-1} \eta_2^{-1} \underbrace{\int \mathcal{N}\left(\theta|\mathbf{m}_N, \mathbf{S}_N\right) \, \mathrm{d}\theta}_{=1} && (28)\\
&= \eta_1^{-1} \eta_2^{-1} && (29)
\end{aligned}
$$

Thus, the denominator cancels out the normalization constants and we derived the posterior $P(\theta|y_{1:N}, \mathbf{x}_{1:N})$ by :

$$
P(\theta|y_{1:N}, \mathbf{x}_{1:N}) = \mathcal{N}\left(\theta|\mathbf{m}_N, \mathbf{S}_N\right), \tag{30}
$$

with

$$
\begin{aligned}
\mathbf{S}_N &= \left(\sigma^{-2} \mathbf{\Phi}^T \mathbf{\Phi} + \mathbf{S}_0^{-1}\right)^{-1} && (31)\\
\mathbf{m}_N &= \mathbf{S}_N \left(\sigma^{-2} \mathbf{\Phi}^T \mathbf{y} + \mathbf{S}_0^{-1} \mathbf{m}_0\right) && (32)
\end{aligned}
$$

That the posterior is of the same form, i.e., a Gaussian, as the prior is a convenient property and the prior is therefore *conjugate* for the likelihood (see also Deisenroth et al., Chap 6.6.1).

Definition 1 *(Conjugate Prior) A prior is* conjugate *for the likelihood function if the posterior is of the same form/type as the prior.*

Conjugacy is a desired property as it allows to update the parameters of the prior given some data in a closed form. We saw that the Gaussian is self-conjugate with known σ^2, but there are several other combinations of likelihood distributions and conjugate priors (see Prince, Chap. 3).

What follows from this property is that we can use the posterior as prior when getting more data, which is not possible when we just have a point estimate of the parameters as with the maximum likelihood or maximum a posterior estimate.

2.2 Derivation of the Posterior Prediction

As we now have the posterior derived, we can move on to the posterior predictions $P(y_*|\mathbf{x}_{1:N}, y_{1:N}, \mathbf{x}_*)$ for an unseen feature vector $\mathbf{x}_*$:

$$
\begin{aligned}
P(y_*|\mathbf{x}_{1:N}, y_{1:N}, \mathbf{x}_*) &= \int P(y_*|\mathbf{x}_*, \theta) P(\theta|\mathbf{x}_{1:N}, y_{1:N}) \, \mathrm{d}\theta && (33)\\
&= \int \mathcal{N}(y_*|\phi(\mathbf{x}_*)\theta, \sigma^2) \mathcal{N}(\theta|\mathbf{m}_N, \mathbf{S}_N) \, \mathrm{d}\theta && (34)
\end{aligned}
$$

Instead of using the change of variables, deriving the product, and then transforming the remaining normalization constants, we refer here to the derivation of Prince given in the solutions [2].

It turns out that (see Deisenroth, Chap. 9.3.4):

$$P(y_*|\mathbf{x}_{1:N}, y_{1:N}, \mathbf{x}_*) = \mathcal{N}(y_*|\phi^T(\mathbf{x}_*)\mathbf{m}_N, \phi^T(\mathbf{x}_*)\mathbf{S}_N\phi(\mathbf{x}_*) + \sigma^2) \tag{35}$$

Equipped with this knowledge, we can now get posteriors for unseen data and by sampling from $P(\theta|\mathbf{x}_{1:N}, y_{1:N})$, we can get different instantiations of a function that fits the data. By visualizing the confidence bounds (as given by the variance of the posterior predictions), we can see where the model is uncertain. With this information, we can potentially gather additional training data in these uncertain regions to improve our model.

[2] http://www0.cs.ucl.ac.uk/external/s.prince/book/AnswerBookletStudents.pdf