Gaussian Process

Eugenio Fella, Theivan Pasupathipillai, Matteo Pedrazzi,
Gaetano Ricucci, Carlo Sgorlon Gaiatto

June 28, 2023

1 Bayesian linear regression

1.1 Linear regression model

Let's consider a linear regression model:

$$\mathbf{y} = \mathbf{X}\mathbf{w} \tag{1}$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix} \quad \text{with} \quad \mathbf{x}_i, \mathbf{w} \in \mathbb{R}^{d+1} \tag{2}$$

The posterior probability of the regression coefficients is:

$$p(\mathbf{w}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})} \tag{3}$$

1.2 Likelihood

Let's assume that each observation y_i is drawn from a Gaussian distribution with unknown mean $\sum_j X_{ij} w_j$ and fixed variance σ_i^2. Therefore the likelihood is:

$$p\left(\mathbf{y} \mid \mathbf{w}\right) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\mathbf{w}, \mathbf{\Sigma}\right) \tag{4}$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix} \tag{5}$$

Result 1.1. *The product of N independent Gaussian distributions is a multivariate Gaussian distribution.*

Proof.

$$\begin{aligned}
\log p\left(\mathbf{y} \mid \mathbf{w}\right) &= \log \prod_{i=1}^{N} \mathcal{N}\left(y_i \mid \sum_j X_{ij} w_j, \sigma_i^2\right) \\
&= \sum_{i=1}^{N} \log \mathcal{N}\left(y_i \mid \sum_j X_{ij} w_j, \sigma_i^2\right) \\
&= -\frac{1}{2}\left[\frac{\left(y_1 - \sum_j X_{1j} w_j\right)^2}{\sigma_1^2} + \frac{\left(y_2 - \sum_j X_{2j} w_j\right)^2}{\sigma_2^2} + \cdots + \frac{\left(y_N - \sum_j X_{Nj} w_j\right)^2}{\sigma_N^2}\right] + const \\
&= -\frac{1}{2}\begin{pmatrix} y_1 - \sum_j X_{1j} w_1 \\ y_2 - \sum_j X_{2j} w_2 \\ \vdots \\ y_N - \sum_j X_{Nj} w_N \end{pmatrix}^T \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_N^2} \end{pmatrix} \begin{pmatrix} y_1 - \sum_j X_{1j} w_1 \\ y_2 - \sum_j X_{2j} w_2 \\ \vdots \\ y_N - \sum_j X_{Nj} w_N \end{pmatrix} + const \\
&= -\frac{1}{2}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)^T \mathbf{\Sigma}^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right) + const \\
&= \log \mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\mathbf{w}, \mathbf{\Sigma}\right)
\end{aligned}$$

□

1.3 Prior

The prior is chosen to be a conjugate prior, namely it is also Gaussian:

$$p\left(\mathbf{w}\right) = \mathcal{N}\left(\mathbf{w} \mid \mu, \mathbf{V_0}\right) \tag{6}$$

1.4 Joint probability

The joint probability is the product of two multivariate Gaussian distributions that can be shown to be again Gaussian:

$$p\left(\mathbf{w}, \mathbf{y}\right) = p(\mathbf{y}|\mathbf{w})p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w}, \mathbf{y} \mid \mu', \mathbf{\Sigma}'\right) \tag{7}$$

where

$$\mu' = \begin{pmatrix} \mu \\ \mathbf{X}\mu \end{pmatrix} \quad \mathbf{\Sigma}' = \begin{pmatrix} \mathbf{V_0} & \mathbf{V_0}\mathbf{X}^T \\ \mathbf{X}\mathbf{V_0} & \mathbf{\Sigma} + \mathbf{X}\mathbf{V_0}\mathbf{X}^T \end{pmatrix} \tag{8}$$

Result 1.2. *The product of two multivariate Gaussian distributions is again Gaussian.*

Proof.

$$\begin{aligned}
\log p\left(\mathbf{w}, \mathbf{y}\right) &= \log p\left(\mathbf{y} \mid \mathbf{w}\right) p\left(\mathbf{w}\right) && (9) \\
&= -\frac{1}{2}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)^T \mathbf{\Sigma}^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right) + -\frac{1}{2}\left(\mathbf{w} - \mu\right)^T \mathbf{V_0}^{-1}\left(\mathbf{w} - \mu\right) + const && (10) \\
&= -\frac{1}{2}\left[\mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{y} - \mathbf{y}^T\mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{w} - \left(\mathbf{X}\mathbf{w}\right)^T \mathbf{\Sigma}^{-1}\mathbf{y} + \left(\mathbf{X}\mathbf{w}\right)^T \mathbf{\Sigma}^{-1}\mathbf{X}\mathbf{w}\right] && (11) \\
&\quad - \frac{1}{2}\left[\mathbf{w}^T\mathbf{V_0}^{-1}\mathbf{w} - \mathbf{w}^T\mathbf{V_0}^{-1}\mu - \mu^T\mathbf{V_0}^{-1}\mathbf{w} + \mu^T\mathbf{V_0}^{-1}\mu\right] + const && (12)
\end{aligned}$$

Let's separate second order terms with respect to the linear ones and collect everything that does not depend on $\mathbf{w}$ or $\mathbf{y}$ into *const*. Moreover note that:

$$(\mathbf{X}\mathbf{w})^T = \mathbf{w}^T\mathbf{X}^T \tag{13}$$

$$\mu^T\mathbf{V_0}^{-1}\mathbf{w} = \left(\mathbf{w}^T\mathbf{V_0}^{-1}\mu\right)^T = \mathbf{w}^T\mathbf{V_0}^{-1}\mu \tag{14}$$

Therefore we have:

$$\log p\left(\mathbf{w},\mathbf{y}\right) = -\frac{1}{2}\mathbf{w}^T\left(\mathbf{V_0}^{-1} + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)\mathbf{w}\ +\frac{1}{2}\mathbf{w}^T\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}\ +\frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\mathbf{w}\ -\frac{1}{2}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}\ +\mathbf{w}^T\mathbf{V_0}^{-1}\mu + const \tag{15}$$

$$= -\frac{1}{2}\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}^T\begin{pmatrix}\mathbf{V_0}^{-1} + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} & -\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\\ -\boldsymbol{\Sigma}^{-1}\mathbf{X} & \boldsymbol{\Sigma}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix} + \begin{pmatrix}\mathbf{w}\\ \mathbf{y}\end{pmatrix}^T\begin{pmatrix}\mathbf{V_0}^{-1}\mu\\ \mathbf{0}\end{pmatrix} + const \tag{16}$$

To determine the covariance matrix, let's define:

$$\boldsymbol{\Sigma}^{'-1} = \begin{pmatrix}\mathbf{V_0}^{-1} + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} & -\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\\ -\boldsymbol{\Sigma}^{-1}\mathbf{X} & \boldsymbol{\Sigma}^{-1}\end{pmatrix} \tag{17}$$

To get $\boldsymbol{\Sigma}^{'}$ we use the following result for the inverse of a partitioned matrix:

$$\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}^{-1} = \begin{pmatrix}\mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1}\\ -\mathbf{D}^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1}\end{pmatrix} \tag{18}$$

where

$$\mathbf{M} = \left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right)^{-1} \tag{19}$$

Doing the (simple) math we get:

$$\boldsymbol{\Sigma}^{'} = \begin{pmatrix}\mathbf{V_0} & \mathbf{V_0}\mathbf{X}^T\\ \mathbf{X}\mathbf{V_0} & \boldsymbol{\Sigma} + \mathbf{X}\mathbf{V_0}\mathbf{X}^T\end{pmatrix} \tag{20}$$

To determine the mean, note that we can write the exponent of a multivariate Gaussian distribution in a general form involving a quadratic term, a linear term and a constant term:

$$-\frac{1}{2}(\mathbf{z} - \mu')^T\boldsymbol{\Sigma}^{'-1}(\mathbf{z} - \mu') = -\frac{1}{2}\mathbf{z}^T\boldsymbol{\Sigma}^{'-1}\mathbf{z} + \mathbf{z}^T\boldsymbol{\Sigma}^{'-1}\mu' + const \tag{21}$$

Thus, comparing with our result we can get the mean μ':

$$\boldsymbol{\Sigma}^{'-1}\mu' = \begin{pmatrix}\mathbf{V_0}^{-1}\mu\\ \mathbf{0}\end{pmatrix} \tag{22}$$

multiplying by $\boldsymbol{\Sigma}^{'}$

$$\mu' = \boldsymbol{\Sigma}^{'}\begin{pmatrix}\mathbf{V_0}^{-1}\mu\\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mu\\ \mathbf{X}\mu\end{pmatrix} \tag{23}$$

$\square$

1.5 Posterior

The posterior can be found by starting with the joint probability and showing that the conditional probability is also Gaussian:

$$p(\mathbf{w} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{w} \mid \mu_{\mathbf{w}|\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{w}|\mathbf{y}}\right) \tag{24}$$

where

$$\mu_{\mathbf{w}|\mathbf{y}} = \mu + \mathbf{V}_0\mathbf{X}^T\left(\mathbf{\Sigma} + \mathbf{X}\mathbf{V}_0\mathbf{X}^T\right)^{-1}(\mathbf{y} - \mathbf{X}\mu) \tag{25}$$

$$\mathbf{\Sigma}_{\mathbf{w}|\mathbf{y}} = \mathbf{V}_0 - \mathbf{V}_0\mathbf{X}^T\left(\mathbf{\Sigma} + \mathbf{X}\mathbf{V}_0\mathbf{X}^T\right)^{-1}\mathbf{X}\mathbf{V}_0 \tag{26}$$

Result 1.3. *The conditional distribution derived from a multivariate Gaussian distribution is again Gaussian.*

Proof. TODO □

2 Bayesian nonparametric linear regression

This approach allows to make predictions at new locations:

$$\mathbf{y}^* = \mathbf{X}^*\mathbf{w} \tag{27}$$

It can be shown that the distribution of predictions given the observation is again Gaussian and independent of the regression coefficients:

$$p(\mathbf{y}^* \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^* \mid \mu_{\mathbf{y}^*|\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{y}^*|\mathbf{y}}\right) \tag{28}$$

where

$$\mu_{\mathbf{y}^*|\mathbf{y}} = \mathbf{X}^*\mu_{\mathbf{y}^*|\mathbf{y}} = \mathbf{X}^*\mu + \mathbf{X}^*\mathbf{V}_0\mathbf{X}^T\left(\mathbf{\Sigma} + \mathbf{X}\mathbf{V}_0\mathbf{X}^T\right)^{-1}(\mathbf{y} - \mathbf{X}\mu) \tag{29}$$

$$\mathbf{\Sigma}_{\mathbf{y}^*|\mathbf{y}} = \mathbf{X}^*\mathbf{\Sigma}_{\mathbf{y}^*|\mathbf{y}}\mathbf{X}^{*T} = \mathbf{X}^*\mathbf{V}_0\mathbf{X}^{*T} - \mathbf{X}^*\mathbf{V}_0\mathbf{X}^T\left(\mathbf{\Sigma} + \mathbf{X}\mathbf{V}_0\mathbf{X}^T\right)^{-1}\mathbf{X}\mathbf{V}_0\mathbf{X}^{*T} \tag{30}$$

Result 2.1. *The distribution of a linear transformation of Gaussian distributed random variable is again Gaussian*

Proof.

$$\mu_{\mathbf{y}^*|\mathbf{y}} = \mathbb{E}\left[\mathbf{y}^*\right] = \mathbb{E}\left[\mathbf{X}^*\mathbf{w}\right] = \mathbf{X}^*\mathbb{E}\left[\mathbf{w}\right] = \mathbf{X}^*\mu_{\mathbf{w}|\mathbf{y}} \tag{31}$$

$$\mathbf{\Sigma}_{\mathbf{y}^*|\mathbf{y}} = \mathbb{E}\left[\left(\mathbf{y}^* - \mathbb{E}\left[\mathbf{y}^*\right]\right)\left(\mathbf{y}^* - \mathbb{E}\left[\mathbf{y}^*\right]\right)^T\right] = \mathbf{X}^*\mathbb{E}\left[\left(\mathbf{w} - \mathbb{E}\left[\mathbf{w}\right]\right)\left(\mathbf{w} - \mathbb{E}\left[\mathbf{w}\right]\right)^T\right]\mathbf{X}^{*T} = \mathbf{X}^*\mathbf{\Sigma}_{\mathbf{w}|\mathbf{y}}\mathbf{X}^{*T} \tag{32}$$

□

2.1 Kernel trick

To increase the expressiveness of the model it is common to use a non linear feature mapping:

$$\mathbf{\Phi} = \phi(\mathbf{X}) \tag{33}$$

The kernel trick emerges from the observation that the inner products $\Phi\mathbf{V_0}\Phi^T$ can be equivalently computed by evaluating the corresponding kernel function k for all pairs to form the matrix $\mathbf{K_{XX}}$:

$$\mathbf{K_{XX}}(i,j) = k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \mathbf{\Phi}\left(\mathbf{x}_i\right)\mathbf{V}_0\mathbf{\Phi}\left(\mathbf{x}_j\right)^T = \left\langle \mathbf{\Phi}\left(\mathbf{x}_i\right), \mathbf{\Phi}\left(\mathbf{x}_j\right)\right\rangle_{\mathbf{V}_0} \tag{34}$$

The kernel trick allows us to specify an intuitive similarity between pairs of points, rather than a feature map Φ, which in practice can be hard to define.

2.2 Gaussian Process

Applying the feature map and the kernel trick describe before, we obtain a Gaussian Process:

$$p(\mathbf{y}^* \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{y}^* \mid \mu_{\mathbf{y}^*|\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{y}^*|\mathbf{y}}\right) \tag{35}$$

where

$$\mu_{\mathbf{y}^*|\mathbf{y}} = \mathbf{\Phi}^* \mu + \mathbf{K}_{\mathbf{X}^*\mathbf{X}} \hat{\mathbf{K}}_{\mathbf{X}\mathbf{X}}^{-1} \left(\mathbf{y} - \mathbf{\Phi}\mu\right) \tag{36}$$

$$\mathbf{\Sigma}_{\mathbf{y}^*|\mathbf{y}} = \mathbf{K}_{\mathbf{X}^*\mathbf{X}^*} - \mathbf{K}_{\mathbf{X}^*\mathbf{X}} \hat{\mathbf{K}}_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{K}_{\mathbf{X}\mathbf{X}^*} \tag{37}$$