Gaussian Process Regression

Uri Shaham

January 21, 2025

1 Preliminary: Multivariate Gaussians

Definition 1.1 (multivariate normal distribution)**.** *A random vector $X \in \mathbb{R}^d$ is said to have a multivariate normal distribution with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix Σ if Σ is positive definite and x has density*

$$p(X; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu) \right).$$

We can partition the d variables to two sets, A and B. In this case we can write $X = \begin{bmatrix} X_A \\ X_B \end{bmatrix}$, $\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}$, $\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$.

Proposition 1.2. *The conditional density $p(X_A|X_B) = \frac{p(X_A, X_B; \mu, \Sigma)}{\int_{X_A} p(X_A, X_B; \mu, \Sigma) dX_B}$ are also multivariate normal:*

$$p(X_A|X_B) \sim \mathcal{N}\left(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(X_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right).$$

Proof.

$$\begin{aligned} p(X_A|X_B) &= \frac{p(X_A, X_B; \mu, \Sigma)}{\int_{X_A} p(X_A, X_B; \mu, \Sigma) dX_B} \\ &= \frac{1}{\int_{X_A} p(X_A, X_B; \mu, \Sigma) dX_B} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu) \right) \\ &= \frac{1}{Z} \exp\left(-\frac{1}{2} \begin{bmatrix} X_A - \mu_A \\ X_B - \mu_B \end{bmatrix}^T \begin{bmatrix} V_{AA} \, V_{AB} \\ V_{BA} \, V_{BB} \end{bmatrix} \begin{bmatrix} X_A - \mu_A \\ X_B - \mu_B \end{bmatrix} \right), \end{aligned} \tag{1}$$

where Z does not depend on X_A, and $\Sigma^{-1} = V = \begin{bmatrix} V_{AA} \, V_{AB} \\ V_{BA} \, V_{BB} \end{bmatrix}$. Observe that

$$\begin{aligned} \begin{bmatrix} X_A - \mu_A \\ X_B - \mu_B \end{bmatrix}^T \begin{bmatrix} V_{AA} \, V_{AB} \\ V_{BA} \, V_{BB} \end{bmatrix} \begin{bmatrix} X_A - \mu_A \\ X_B - \mu_B \end{bmatrix} &= (X_A - \mu_A)^T V_{AA} (X_A - \mu_A) \\ &+ (X_A - \mu_A)^T V_{AB} (X_B - \mu_B) \\ &+ (X_B - \mu_B)^T V_{BA} (X_A - \mu_A) \\ &+ (X_B - \mu_B)^T V_{BB} (X_B - \mu_B). \end{aligned}$$

Retaining only terms depending on X_A, and using the fact that $V_{AB} = V_{BA}^T$ we can thus have

$$p(X_A|X_B) \propto \exp\left(-\frac{1}{2}\left[X_A^T V_{AA} X_A - 2X_A^T V_{AA}\mu_A + 2X_A^T V_{AB} X_B\right]\right),$$

where the term inside the exponential is $X_A^T V_{AA} X_A - 2X_A^T \left(V_{AA}\mu_A - V_{AB}X_B\right)$. Completing the squares[1], we can write this as

$$\left(X_A - (\mu_A - V_{AA}^{-1}V_{AB}X_B)\right)^T V_{AA}\left(X_A - (\mu_A - V_{AA}^{-1}V_{AB}X_B)\right) + c,$$

where c is a constant not depending on X_A. From this, we deduce that $p(X_A|X_B)$ is normal with mean $\mu = \mu_A - V_{AA}^{-1}V_{AB}X_B$ and covariance V_{AA}^{-1}. Finally, we recall the form of the inverse of a block matrix to have $V_{AA} = \left(\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)^{-1}$, and $V_{BA} = -\left(\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)^{-1} - \Sigma_{AB}\Sigma_{BB}^{-1}$. □

2 Gaussian Processes

Gaussian processes are extension of multivariate Gaussians from vectors to functions.

Definition 2.1 (Gaussian process)**.** *A GP with mean function $m(\cdot)$ and covariance function $k(\cdot,\cdot)$ is a stochastic process $\{Z_t : t \in \mathcal{T}\}$ such that for every finite collection $t_1, \ldots, t_n$ of indices, the vector $(Z_{t_1}, \ldots, Z_{t_n})^T$ has a multivariate normal distribution with mean vector $\mu = (m(Z_{t_1}), \ldots, m(Z_{t_n}))^T$ and covariance matrix K such that $K_{ij} = k(Z_{t_i}, Z_{t_j})$. We*

Since the covariance function has to be positive definite, it makes sense that k will be a kernel function. A typical choice for k is the RBF function

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{\sigma^2}\right).$$

When we say that a function f is a sample drawn from a GP prior, we can think of f as a sample from a infinite dimensional multivariate normal vector, where each entry corresponds to an index $t \in \mathcal{T}$. That is, $f = \{f(x_t) : t \in \mathcal{T}\}$.

3 Gaussian Process Regression

GPR is a popular tool to quantify prediction uncertainty. Let $\{(x_i, y_i)\},\ i = 1, \ldots n$ be a training set, drawn from some data distribution $\mathcal{D}$. where $x_i \in \mathbb{R}^d$, and $y_i \in \mathbb{R}$. We model the data by $y_i = f(x_i) + \epsilon_i$, where $f : \mathbb{R}^d \to \mathbb{R}$ is some function, drawn from a GP prior with zero mean and covariance function k, and the ϵ_i's are iid samples from zero mean normal distribution with variance σ^2. Let $\{(x_j^*)\},\ j = 1, \ldots m$ be a test set, also drawn from the x-marginal distribution induced from $\mathcal{D}$. In vector form we can write

$$\vec{y} = \vec{f} + \vec{\epsilon},$$

and

$$\vec{y^*} = \vec{f^*} + \vec{\epsilon^*},$$

where $\vec{y} = (y_1, \ldots, y_n)^T$, $\vec{f} = (f(x_1), \ldots, f(x_n))^T$, and so on. Similarly, we write $X = (x_1^T, \ldots x_n^T)^T$, i.e., an $n \times d$ matrix.

[1] https://en.wikipedia.org/wiki/Completing_the_square

3.1 Prediction

Since f is a sample from a Gaussian process prior with zero mean vector and covariance function k, it follows that given X, X^*

$$\begin{bmatrix} \vec{f} \\ \vec{f}^* \end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} k(X,X) & k(X,X^*) \\ k(X^*,X) & k(X^*,X^*) \end{bmatrix}, \right)$$

where $k(X,X^*)_{ij} = k(x_i, x_j^*)$, and so on. Since both $\vec{f}$ and $\vec{epsilon}$ are Gaussians, it follows that so is $\vec{y}$, i.e., given X, X^*

$$\begin{bmatrix} \vec{Y} \\ \vec{Y}^* \end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} k(X,X) + \sigma^2 I & k(X,X^*) \\ k(X^*,X) & k(X^*,X^*) + \sigma^2 I \end{bmatrix}. \right)$$

Finally, we are interested in the predictive distribution

$$p(\vec{Y}^* | X, X^*, \vec{y}).$$

recalling proposition 1.2, this distribution is multivariate Gaussian

$$p(\vec{Y}^* | X, X^*, y^*) = \mathcal{N}\left(\mu^*, \Sigma^*\right), \tag{2}$$

with

$$\mu^* = k(X,X^*)(k(X,X) + \sigma^2 I)^{-1}\vec{y},$$

and

$$\Sigma^* = k(X^*,X^*) + \sigma^2 I) - k(X^*,X)((k(X,X) + \sigma^2 I)^{-1} k(X,X^*).$$

And that's it :)

In particular, this gives us a measure of uncertainty in the prediction of y_j^*, which is Σ_{jj}^*, as

$$Y_j^* \sim \mathcal{N}\left(\mu_j^*, \Sigma_{jj}^*\right).$$

Remark 3.1. *The noise variance σ^2 can be estimated using a validation set, by a grid search, where we look for the value that reduces the validation error the most.*

4 Application: Bayesian optimization

The goal to optimize the hyperparameters of a machine learning model (say, a neural network).

- Clearly, these cannot be optimized using gradient-based optimization.
- However, given a setting of hyperparameters we can train the model and measure its validation loss.
- We consider the Validation loss as a GP, defined over the space of all configurations of hyperparameters. In this case X will correspond to a configuration of hyperparameters and our kernel function k will quantify the similarity between two different configurations.
- Once we have a few loss measurements at few random configurations of hyperparameters, we can use equation 2, taking X to be configurations of hyperparameters and Y to be validation loss.
- Given a future configuration to test, we can compute the expected improvement, or the probability of improvement over the current best configuration, and use this as a criterion for selection of the figure configuration to try.

- Once we try a new configuration, we repeat the process (with updated conditional mean and covariance).
- Note that since different hyperparameters have different scales, it's useful to standardize the scales of all hyperparameters.