Manifold Learning for Signal and Visual Processing
Lecture 10: Introduction to Gaussian Processes

Radu Horaud
INRIA Grenoble Rhone-Alpes, France
Radu.Horaud@inria.fr
https://team.inria.fr/perception
Outline of This Lecture

• Back to linear regression.
• Bayesian linear regression.
• Gaussian process problem statement.
• Gaussian process for regression.
Material for This Lecture

- More involved readings:
Linear Regression

- The task is to estimate the parameters \( \mathbf{w} = (w_0, \ldots, w_{M-1})^\top \) in:

\[
y(x, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x) = \mathbf{w}^\top \phi(x)
\]

- from a training dataset of \( N \) observations \( \{x_n\} \) together with corresponding target values \( \{t_n\} \).

- The target value is given by \( t = y(x, \mathbf{w}) + \epsilon \) where \( \epsilon \) is a zero-mean Gaussian variable with precision \( \beta \). Thus we can write:

\[
p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})
\]
Maximum Likelihood Estimation

- The log-likelihood function is, with $t = (t_1, \ldots, t_n)$:

$$\ln p(t|X, w, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n|w^\top \phi(x_n), \beta^{-1})$$

- MLE solution:

$$w_{ML} = (\Phi^\top \Phi)^{-1} \Phi^\top t$$

- and the variance (inverse of precision):

$$\beta_{ML}^{-1} = \frac{1}{N} \sum_{n=1}^{N} (t_n - w_{ML}^\top \phi(x_n))^2$$
Bayesian Linear Regression

- Prior distribution of the model parameters:

\[ p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \]

- The posterior distribution is proportional to the product of the likelihood function \( p(\mathbf{t}|\mathbf{w}) \) and the prior \( p(\mathbf{w}) \). This results in (see chapter 2 of Bishop, eq. (2.116)):

\[ p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \]

- with:

\[
\begin{align*}
\mathbf{m}_N &= \beta \mathbf{S}_N \Phi^\top \mathbf{t} \\
\mathbf{S}_{N}^{-1} &= \alpha \mathbf{I} + \beta \Phi^\top \Phi^\top.
\end{align*}
\]
Maximization of the Posterior Distribution

\[
\ln p(t|w) = -\frac{\beta}{2} \sum_{n=1}^{N} (t_n - w^\top \phi(x_n))^2 - \frac{\alpha}{2} w^\top w + \text{const.}
\]

- This is equivalent to a regularized least-square optimization problem with respect to \( w \) and with a regularization coefficient \( \lambda = \alpha/\beta \) (see Bishop section 3.14).
Predictive Distribution

- Predict a value for $t$ for a new value of $x$ given the model parameters $w$ (I simplified the Bishop notations):

$$p(t) \propto \int p(t|w)p(w)dw$$

- Using the previous formulae we obtain the \textit{predictive distribution} $p(t|w)$:

$$p(t|x, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t, m_N^\top \phi(x), \sigma^2_N(x))$$

$$\sigma^2_N(x) = \frac{1}{\beta} + \phi(x)^\top S_N \phi(x).$$
Linear Regression Again

\[ y(x) = w^\top \phi(x) \]

Let \( y = \{y_n\} \) with \( y_n = y(x_n) \) and let \( \Phi \) be the following \( N \times M \) matrix:

\[
\Phi = 
\begin{bmatrix}
\phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_{M-1}(x_1) \\
\vdots & \vdots & \ddots & \vdots \\
\phi_0(x_N) & \phi_1(x_N) & \ldots & \phi_{M-1}(x_N)
\end{bmatrix}
\]

Altogether this can be written:

\[ y = \Phi w \]
**Probability Distribution of \( y \)**

- We seek the joint probability distribution of \( y = \{ y_1, \ldots, y_N \} \) using the prior \( p(w) = \mathcal{N}(w|0, \alpha^{-1}I) \).
- \( y \) is a linear combination of Gaussian variables, it is Gaussian itself, with:

\[
E[y] = \Phi E[w] = 0
\]

\[
cov[y] = E[yy^\top] = \Phi E[ww^\top] \Phi^\top = \frac{1}{\alpha} \Phi \Phi^\top = K
\]

- The matrix \( K \) is a Gramm (or kernel) matrix with elements given by:

\[
K_{nm} = \frac{1}{\alpha} \phi(x_n)^\top \phi(x_m) = k(x_n, x_m)
\]
A Gaussian process (GP) is defined as probability distribution over functions \( y(x) \) such that the set of values \( y_1, \ldots, y_N \) evaluated at an arbitrary set of points \( x_1, \ldots, x_N \), jointly have a Gaussian distribution.

When \( x \in \mathbb{R}^2 \) this is known as a Gaussian random field.

A stochastic process \( y(x) \) is specified by the joint probability distribution for a finite set of values in a consistent manner.

Key point about stochastic GP: the joint distribution \( p(y_1, \ldots, y_N) \) is specified completely by the second-order statistics (mean and covariance).
GP for Regression

- Let’s consider again that the observed target values have Gaussian noise, namely $t_n = y_n + \epsilon_n$, or:

$$p(t_n|y_n) = \mathcal{N}(t_n|y_n, \beta^{-1})$$

- The joint distribution of the target values:

$$p(t|y) = \mathcal{N}(t|y, \beta^{-1}I)$$

- From above, we can write the distribution of $y$:

$$p(y) = \mathcal{N}(y|0, K)$$

where $K$ is a kernel matrix. The kernel must be chosen such as to express the fact that if $x_n$ and $x_m$ are similar the corresponding values of $y(x_n)$ and $y(x_m)$ will be strongly correlated.
The marginal distribution $p(t)$, conditioned on the observed input values $\{x_n\}$, can be found by integrating over $y$ and using the results of chapter 2 (Bishop):

$$p(t) = \int p(t|y)p(y)dy = \mathcal{N}(t|0, C)$$

with:

$$C = K + \beta^{-1}I$$
A Gaussian Process Example

- A “widely” used kernel function:

\[ k(x_n, x_m) = \theta_0 \exp \left( -\frac{\theta_1}{2} \|x_n - x_m\|^2 \right) + \theta_2 + \theta_3 x_n^\top x_m \]

- \( \theta = (\theta_0, \theta_1, \theta_2, \theta_3) \) are the hyper-parameters of the GP model.
- We found an expression for \( p(t) \) but it would be more correct to write \( p(t|\theta) \).
The Predictive Distribution

- Let $x_{N+1}$ be a new input vector and we want to predict the target $t_{N+1}$. We need to compute the predictive distribution $p(t_{N+1} | t_N, X_{N+1})$ with $X_{N+1} = (x_1, \ldots, x_N, x_{N+1})$, for convenience we drop $X$.

- Hence the predictive distribution is written $p(t_{N+1} | t_N)$. It can be obtained from the joint distribution:

$$p(t_{N+1}) = \mathcal{N}(t_{N+1} | 0, C_{N+1})$$

- The covariance matrix can be decomposed such as to use the covariance that was estimated from the training data, namely:

$$C_{N+1} = \begin{bmatrix}
K_N + \beta^{-1}I_N \\
\vdots \\
k(x_1, x_{N+1}) & \cdots & k(x_N, x_{N+1}) \\
\vdots \\
k(x_{N+1}, x_{N+1}) + \beta^{-1}
\end{bmatrix}$$
Using various results (Bishop section 2, eqs. (2.81) and (2.82)) we see that the predictive (conditional) distribution is Gaussian with mean and covariance:

\[ p(t_{N+1}|t_N) = \mathcal{N}(t_{N+1}| k^\top C_N^{-1} t_N, c - k^\top C_N^{-1} k) \]
The training involves the estimation of the hyper-parameters $\theta$ by maximizing the log-likelihood function $\ln p(t_N|\theta)$ which is a non-convex problem.

It is possible to introduce a prior $p(\theta)$ and maximize the log-posterior using gradient-based methods. This implies marginalization, which is intractable in the general case.

The covariance matrix $C_N$ of size $N \times N$ must be inverted.

There is no general methodology to select the kernel function!