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

- C. M. Bishop (2006). Pattern Recognition and Machine Learning. Chapters 3 and 6 (section 6.4).
- More involved readings:

  Rasmussen and Williams (2006). Gaussian Processes for Machine Learning.

Linear Regression

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

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

- from a training dataset of N observations $\{\boldsymbol{x}_n\}$ together with corresponding target values $\{t_n\}$.
- The target value is given by $t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$ where ϵ is a zero-mean Gaussian variable with precision β. Thus we can write:

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

Maximum Likelihood Estimation

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

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

- MLE solution:

$$\boldsymbol{w}_{\mathsf{ML}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{t}$$

- and the variance (inverse of precision):

$$\beta_{\mathsf{ML}}^{-1} = \frac{1}{N} \sum_{n=1}^{N} (t_n - \boldsymbol{w}_{\mathsf{ML}}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2$$

Bayesian Linear Regression

- Prior distribution of the model parameters:

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

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

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

- with:

$$\begin{aligned}\boldsymbol{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^\top \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}^\top.\end{aligned}$$

Maximization of the Posterior Distribution

$$\ln p(\mathbf{t}|\boldsymbol{w}) = -\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n))^2 - \frac{\alpha}{2}\boldsymbol{w}^\top \boldsymbol{w} + \text{const.}$$

- This is equivalent to a regularized least-square optimization problem with respect to $\boldsymbol{w}$ 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 $\boldsymbol{x}$ given the model parameters $\boldsymbol{w}$ (I simplified the Bishop notations):

$$p(t) \propto \int p(t|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

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

$$p(t|\boldsymbol{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t, \boldsymbol{m}_N^\top \boldsymbol{\phi}(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x}))$$
$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^\top \mathbf{S}_N \boldsymbol{\phi}(\boldsymbol{x}).$$

Linear Regression Again

$$y(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

- Let $\mathbf{y} = \{y_n\}$ with $y_n = y(\boldsymbol{x}_n)$ and let $\boldsymbol{\Phi}$ be the following $N \times M$ matrix:

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \dots & \phi_{M-1}(\boldsymbol{x}_1) \\ \vdots & \vdots & & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \dots & \phi_{M-1}(\boldsymbol{x}_N) \end{bmatrix}$$

- Altogether this can be written:

$$\mathbf{y} = \boldsymbol{\Phi} \boldsymbol{w}$$

Probability Distribution of $\mathbf{y}$

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

$$E[\mathbf{y}] = \boldsymbol{\Phi} E[\boldsymbol{w}] = \mathbf{0}$$

$$\text{cov}[\mathbf{y}] = E[\mathbf{y}\mathbf{y}^\top] = \boldsymbol{\Phi} E[\boldsymbol{w}\boldsymbol{w}^\top]\boldsymbol{\Phi}^\top = \frac{1}{\alpha}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \mathbf{K}$$

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

$$K_{nm} = \frac{1}{\alpha}\boldsymbol{\phi}(\boldsymbol{x}_n)^\top \boldsymbol{\phi}(\boldsymbol{x}_m) = k(\boldsymbol{x}_n, \boldsymbol{x}_m)$$

Gaussian Process Definition (Bishop)

- A Gaussian process (GP) is defined as probability distribution over functions $y(\boldsymbol{x})$ such that the set of values $y_1, \ldots, y_N$ evaluated at an arbitrary set of points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, jointly have a Gaussian distribution.
- When $\boldsymbol{x} \in \mathbb{R}^2$ this is known as a Gaussian random field.
- A stochastic process $y(\boldsymbol{x})$ 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:

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

- From above, we can write the distribution of $\mathbf{y}$:

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

where $\mathbf{K}$ is a kernel matrix. The kernel must be chosen such as to express the fact that if $\boldsymbol{x}_n$ and $\boldsymbol{x}_m$ are similar the corresponding values of $y(\boldsymbol{x}_n)$ and $y(\boldsymbol{x}_m)$ will be strongly correlated.

GP for Regression

- The marginal distribution $p(\mathbf{t})$, conditioned on the observed input values $\{\boldsymbol{x}_n\}$, can be found by integrating over $\mathbf{y}$ and using the results of chapter 2 (Bishop):

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

- with:

$$\mathbf{C} = \mathbf{K} + \beta^{-1}\mathbf{I}$$

A Gaussian Process Example

- A "widely" used kernel function:

$$k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \theta_0 \exp\left(-\frac{\theta_1}{2}\|\boldsymbol{x}_n - \boldsymbol{x}_m\|^2\right) + \theta_2 + \theta_3 \boldsymbol{x}_n^\top \boldsymbol{x}_m$$

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

The Predictive Distribution

- Let $\boldsymbol{x}_{N+1}$ be a new input vector and we want to predict the target t_{N+1}. We need to compute the predictive distribution $p(t_{N+1}|\mathbf{t}_N, \mathbf{X}_{N+1})$ with $\mathbf{X}_{N+1} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{x}_{N+1})$, for convenience we drop $\mathbf{X}$.
- Hence the predictive distribution is written $p(t_{N+1}|\mathbf{t}_N)$. It can be obtained from the joint distribution:

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

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

$$\mathbf{C}_{N+1} = \begin{bmatrix} \mathbf{K}_N + \beta^{-1}\mathbf{I}_N & \begin{matrix} k(\boldsymbol{x}_1, \boldsymbol{x}_{N+1}) \\ \vdots \\ k(\boldsymbol{x}_N, \boldsymbol{x}_{N+1}) \end{matrix} \\ \underbrace{\begin{matrix} k(\boldsymbol{x}_1, \boldsymbol{x}_{N+1}) & \ldots & k(\boldsymbol{x}_N, \boldsymbol{x}_{N+1}) \end{matrix}}_{\boldsymbol{k}^\top} & \underbrace{k(\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+1}) + \beta^{-1}}_{c} \end{bmatrix}$$

The Predictive Distribution

- 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:

$$p(t_{N+1}|\mathbf{t}_N) = \mathcal{N}(t_{N+1}|\underbrace{\boldsymbol{k}^\top \mathbf{C}_N^{-1}\mathbf{t}_N}_{\text{mean}}, \underbrace{c - \boldsymbol{k}^\top \mathbf{C}_N^{-1}\boldsymbol{k}}_{\text{cov}})$$

Remarks

- The training involves the estimation of the hyper-parameters $\boldsymbol{\theta}$ by maximizing the log-likelihood function $\ln p(\mathbf{t}_N|\boldsymbol{\theta}$ which is a non-convex problem.
- It is possible to introduce a prior $p(\boldsymbol{\theta})$ and maximize the log-posterior using gradient-based methods. This implies marginalization, which is intractable in the general case.
- The covariance matrix $\mathbf{C}_N$ of size $N \times N$ must be inverted.
- There is no general methodology to select the kernel function!