Machine Learning 1
Lecture 12.3 - Kernel Methods
Gaussian Processes - Definition

Erik Bekkers

(Bishop 6.4.1)
Gaussian Processes

- **Definition (Gaussian Process):**

  A Gaussian process is a collection of random variables, any finite number of which is jointly Gaussian distributed

- **Or put differently (functional viewpoint):**
  - Gaussian processes represent distributions over random functions.
  - The function evaluated at any specific input \( \mathbf{x} \) is a random variable \( f(\mathbf{x}) \), with

\[
\mathbb{E}[f(\mathbf{x})] = m(\mathbf{x})
\]

\[
\text{cov}(f(\mathbf{x}), f(\mathbf{x}')) = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x})) (f(\mathbf{x}') - m(\mathbf{x}'))] = k(\mathbf{x}, \mathbf{x}')
\]
Functional Viewpoint, why is this a GP?

- Take any finite set \( \{x_1, \ldots, x_N\} \) with corresponding random variables \( \{f(x_1), \ldots, f(x_N)\} \) then

\[
p\left( \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{bmatrix} \right) = \mathcal{N} \left( \begin{bmatrix} m(x_1) \\ \vdots \\ m(x_N) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \ldots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \ldots & k(x_N, x_N) \end{bmatrix} \right)
\]

- Consistency requirement: any subset of \( \{f(x_1), \ldots, f(x_N)\} \) should also be Gaussian distributed.

- But that works out because:

\[
p\left( \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \right) = \mathcal{N} \left( \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \right) \quad \Rightarrow \quad p(f_1) = \mathcal{N} \left( m_1, K_{11} \right)
\]
Functions as vectors

- Think of a function $f(\cdot)$ drawn from a $\textit{GP}$ as an extremely high-dimensional vector drawn from an extremely high-dimensional multivariate Gaussian distribution.

- Each dimension corresponds to an element $\mathbf{x} \in \mathbb{R}^n$.

- Each entry of the vector is $f(\mathbf{x})$ for a particular $\mathbf{x} \in \mathbb{R}^n$.

- How do you sample from a $\textit{GP}$:
  - Sample input points $\mathbf{x} \in \mathbb{R}^n$.
  - Construct the Gram matrix $\mathbf{K}$ for all sampled $\mathbf{x}$.

\[
\begin{bmatrix}
  f(x_1) \\
  \vdots \\
  f(x_N)
\end{bmatrix} \sim \mathcal{N}
\begin{bmatrix}
  m(x_1) \\
  \vdots \\
  m(x_N)
\end{bmatrix},
\begin{bmatrix}
  k(x_1, x_1) & \cdots & k(x_1, x_N) \\
  \vdots & \ddots & \vdots \\
  k(x_N, x_1) & \cdots & k(x_N, x_N)
\end{bmatrix}
\]
Example: Bayesian Linear Regression

- Bayesian linear models:
  \[ f(x) = \phi(x)^T w \]

- Prior on \( w \):
  \[ p(w) = \mathcal{N}(w | 0, \Sigma_p) \]

- Then \( f(x) \) is a Gaussian process
  \[
  \mathbb{E}[f(x)] = \phi(x)^T \mathbb{E}[w] = 0 = m(x)
  \]
  \[
  \text{cov}(f(x), f(x')) = \mathbb{E}[f(x)f(x')] = \phi(x)^T \mathbb{E}[ww^T] \phi(x')
  \]
  \[= \phi(x)^T \Sigma_p \phi(x') = k(x, x') \]

- Thus \( f(x_1), \ldots, f(x_N) \) for any \( N \) are jointly Gaussian!
  \[ \implies f(x) \text{ is distributed according to GP with kernel } \]

\( \Sigma_p \) is the prior covariance matrix.