10-708: Probabilistic Graphical Models, Spring 2015

21 : Advanced Gaussian Processes

Lecturer: Eric P. Xing *Scribes: Konstantin Genin, Yutong Zheng*

1 Gaussian Process Inference

A *Gaussian process* (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. This means that a GP can be completely specified by its mean function $m(\cdot)$ and covariance function $k(\cdot,\cdot)$. We say that $\mathbf{f}(x) \sim \mathcal{GP}(m(x), k(x,x')$ iff for all $N \in \mathbb{N}$, $(\mathbf{f}(x_1),\ldots,\mathbf{f}(x_N)) \sim \mathcal{N}(\mathbf{f}\mu, K)$ with $\mathbf{f}\mu_i = m(x_i)$ and $K_{ij} = \text{cov}(\mathbf{f}(x_i), \mathbf{f}(x_j)) = k(x_i, x_j)$.

Assuming that we have access only to noisy function values, i.e. that $y = \mathbf{f}(\mathbf{x}) + \epsilon$, and that ϵ is Gaussian with variance σ_n^2 the prior on the noisy data is

$$\text{cov}(\mathbf{f}y) = K(X,X) + \sigma_n^2 I.$$

We can write the joint distribution of the observed target values $\mathbf{f}y$ and function values at test points X_* as

$$\begin{bmatrix} \mathbf{f}y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K_\theta(X,X) + \sigma^2 I & K_\theta(X,X_*) \\ K_\theta(K_*,X) & K_\theta(X_*,X_*) \end{bmatrix}\right)$$

deriving the conditional distribution using standard facts about the conditional Gaussian, we arrive at the key predictive equations for Gaussian process regression:

$$\begin{aligned} \mathbf{f}_*|X,\mathbf{y},X_* &\sim \mathcal{N}(\overline{\mathbf{f}_*}, \text{cov}(\mathbf{f}_*)), \text{where} \\ \overline{\mathbf{f}_*} &:= K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1}\mathbf{y}, \\ \text{cov}(\mathbf{f}_*) &= K(X_*,X_*) - K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1}K(X,X_*). \end{aligned}$$

We can also marginalize over the function values $\mathbf{f}$ to obtain the marginal likelihood:

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f},X)p(\mathbf{f}|X)d\mathbf{f}.$$

Performing the integration, we obtain the log marginal likelihood:

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi.$$

2 Kernel functions.

Informally, a kernel function k furnishes a notion of similarity between points. Oftentimes, far away points are less similar than nearby points. So, $K_{ij} = \langle\phi(x_i), \phi(x_j)\rangle$ characterizes the "overlap" between features. All

linear basis function models $f(x) = w^T\phi(x)$ with a Gaussian prior over the weight vector: $p(w) = \mathcal{N}(0, \Sigma_w)$ correspond to Gaussian processes with the kernel function $k(x, x') = \phi(x)^T\Sigma_w\phi(x')$. This gives us a canonical way of translating between the weight space view, and the function space view. Some popular kernels include:

$$
\begin{aligned}
k_{SE}(\mathbf{x}, \mathbf{x}') &= \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^2}{2\ell^2}\right) \\
k_{MA}(\mathbf{x}, \mathbf{x}') &= a\left(1 + \frac{\sqrt{3}(\mathbf{x}-\mathbf{x}')}{\ell}\right)\exp\left(-\frac{\sqrt{3}(\mathbf{x}-\mathbf{x}')}{\ell}\right) \\
k_{RQ}(\mathbf{x}, \mathbf{x}') &= \left(1 + \frac{(\mathbf{x}-\mathbf{x}')^2}{2\alpha\ell^2}\right)^{-\alpha} \\
k_{PE}(\mathbf{x}, \mathbf{x}') &= \exp\left(-2\sin^2(\pi(\mathbf{x}-\mathbf{x}')\omega/\ell^2\right)
\end{aligned}
$$

Recall that a positive semi-definite matrix is any $n \times n$ matrix K that satisfies $\mathbf{v}^T K\mathbf{v} \geq 0$ for all vectors $\mathbf{v} \in \mathbb{R}^n$. Recall also that the covariance matrix of a multivariate probability distribution is always a PSD matrix, and that conversely, every PSD matrix is the covariance matrix of some multivariate distribution. For any kernel function $k(\cdot, \cdot)$, and finite set of input points $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$, the *Gram matrix* is the matrix who entries are given by $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. We say that a kernel is PSD if any choice of input points gives rise to a PSD Gram matrix, i.e. if its Gram matrix is a covariance matrix. If this is the case, we say that the kernel function is *valid* for a GP.

A *stationary* kernel function is invariant under translations of the input space, i.e. it is a function of $(\mathbf{x}-\mathbf{x}')$. If it is a function of only $|\mathbf{x}-\mathbf{x}'|$, it is called *isotropic* — it is invariant to all rigid motions. Bochner's theorem characterizes an interesting class of kernel functions:

Theorem 1 (Bochner's theorem) *A complex-valued function on $\mathbb{R}^D$ is the covariance function of a weakly stationary[1], mean square continuous[2] complex-valued random process iff it can be represented as*

$$
k(\mathbf{x}-\mathbf{x}') = \int_{\mathbb{R}^D} \exp\left(2\pi i\mathbf{s}(\mathbf{x}-\mathbf{x}')\right) d\mu(\mathbf{s}),
$$

where μ is a positive, finite measure.

In the case that the spectral density $S(\mathbf{s})$ exists, the covariance function and the spectral density are Fourier duals, by the Wiener-Khintchine theorem:

$$
k(\mathbf{x}-\mathbf{x}') = \int S(\mathbf{s})\exp\left(2\pi i\mathbf{s}(\mathbf{x}-\mathbf{x}')\right) d\mathbf{s}, \qquad S(\mathbf{s}) = \int k(\mathbf{x}-\mathbf{x}')\exp\left(2\pi i\mathbf{s}(\mathbf{x}-\mathbf{x}')\right) d(\mathbf{x}-\mathbf{x}').
$$

2.1 The Squared Exponential Kernel

We show that the squared-exponential kernel, arises as the function-space expression of a radial basis function regression model with infinitely many basis functions. Recall that a linear basis function model given by:

$$
\begin{aligned}
f(x, \mathbf{w}) &= \mathbf{w}^T\phi(x), \\
p(\mathbf{w}) &= \mathcal{N}(0, \Sigma_w),
\end{aligned}
$$

[1] A GP is weakly stationary if its kernel function is stationary, and it has constant mean

[2] A GP is mean-square continuous iff for all $\mathbf{x}, \mathbf{x}'$, $\lim_{\mathbf{x}\to\mathbf{x}'} \mathbb{E}[|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}')|^2] = 0$.

gives rise to the kernel function $k(x_i, x_j) = \phi(x_i)^T \Sigma_w \phi(x_j)$. We start with the finite model:

$$\begin{aligned} f(x) &= \sum_{i=1}^{J} w_i \phi_i(x), \\ w_i &\sim \mathcal{N}\left(0, \frac{\sigma^2}{J}\right), \\ \phi_i(x) &= \exp\left(-\frac{(x - c_i)^2}{2\ell^2}\right). \end{aligned}$$

This gives rise to the kernel function: $k(x, x') = \frac{\sigma^2}{J} \sum_{i=1}^{J} \phi_i(x)\phi_i(x')$. Letting $c_{i+1} - c_i = \frac{1}{J}$ and $J \to \infty$ the kernel becomes a Riemann sum:

$$k(x, x') = \lim_{J \to \infty} \frac{\sigma^2}{J} \sum_{i=1}^{J} \phi_i(x)\phi_i(x') = \int_{c_o}^{c_\infty} \phi_c(x)\phi_c(x')dc.$$

By setting $c_0 = -\infty$ and $C_\infty = \infty$, we spread infinitely many basis functions across the real line, each a distance $\Delta_c \to 0$ apart:

$$\begin{aligned} k(x, x') &= \int_{-\infty}^{\infty} \exp\left(-\frac{(x - c)^2}{2\ell^2}\right) \exp\left(-\frac{(x' - c)^2}{2\ell^2}\right) dc, \\ &= \sqrt{\pi}\ell\sigma^2 \exp\left(-\frac{(x - x')^2}{2(\sqrt{2}\ell)^2}\right), \end{aligned}$$

which we recognize as a squared exponential covariance function with a $\sqrt{2}$ times longer length-scale. What is remarkable is that we have turned a model with infinitely many basis functions into an GP with a particular kind of kernel. This is an instance of the *kernel trick* – replacing inner products of basis functions with kernels. Note that functions drawn from a GP with an RBF kernel are infinitely differentiable. For this reason, Stein accuses it of being overly smooth and unrealistic for modeling many physical processes. Nevertheless, it is the most widely used kernel in the field.

2.2 The Polynomial Kernel

The linear model

$$\begin{aligned} f(x, w) &= \mathbf{w}^T x + b, \\ p(w) &= \mathcal{N}(0, \alpha^2 I), p(b) \qquad\qquad = \mathcal{N}(0, \beta^2), \end{aligned}$$

corresponds to a Gaussian process with the kernel:

$$k_{\text{LIN}}(x, x') = \alpha^2 x^T x + \beta^2.$$

Samples from a GP with this kernel will be straight lines. Recalling that a product of valid kernels is a valid kernel, the product of two linear kernels is a quadratic kernel, giving rise to quadratic functions. This can be generalized to the polynomial kernel $k_{POL}(x, x') = (\alpha^2 x^t x + \beta^2)^p$.

2.3 The Rational Quadratic Kernel

The rational quadratic kernel allows us to model data varying at multiple scales. Letting $r = ||x - x'||$, we set $k(r) = \int \exp(-\frac{r^2}{2\ell^2})p(\ell)d\ell$. If we let $p(\ell)$ take a Gamma density, we derive the rational quadratic kernel as: $k_{\text{RQ}}(r) = (1 + \frac{r^2}{2\alpha\ell'^2})^{-\alpha}$.

2.4 Gibbs Kernel

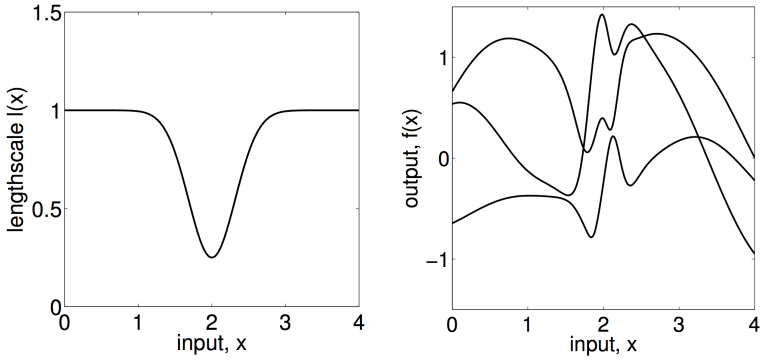
Recall the RBF kernel

$$k_{RBF}(\mathbf{x}, \mathbf{x}') = a^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2l^2}\right)$$

If we want to make the length-scale of l dependent on the feature value, we can change l to a set of $l_p(x)$

$$k_{Gibbs}(\mathbf{x}, \mathbf{x}') = \prod_{p=1}^{P} \left(\frac{2l_p(\mathbf{x})l_p(\mathbf{x}')}{l_p^2(\mathbf{x}) + l_p^2(\mathbf{x}')}\right)^{\frac{1}{2}} \exp\left(-\sum_{p=1}^{P} \frac{(x_p - x_p')^2}{l_p^2(\mathbf{x}) + l_p^2(\mathbf{x}')}\right)$$

where x_p is the p^{th} component of $\mathbf{x}$



(a) (b)

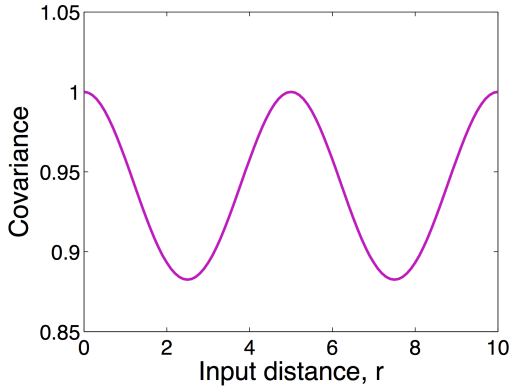
In this figure, the length-scale $l(x)$ is designed to be small around $x = 2$ (a). So the output around $x = 2$ is less smoother.

2.5 Periodic Kernel

For the RBF kernel, if we transfer the inputs through a vector-value function $\mathbf{u}(\mathbf{x}) = (\cos(x), \sin(x))$, the kernel $k_{RBF}(\mathbf{x}, \mathbf{x}') \to k_{RBF}(\mathbf{u}(\mathbf{x}), \mathbf{u}(\mathbf{x}'))$. We have the periodic kernel

$$k_{PER}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{2\sin^2\left(\frac{\mathbf{x}-\mathbf{x}'}{2}\right)}{l^2}\right)$$

An example of periodic kernel is show above. At a certain interval, the input data has a higher covariance with each other.



2.6 Non-Stationary Kernels

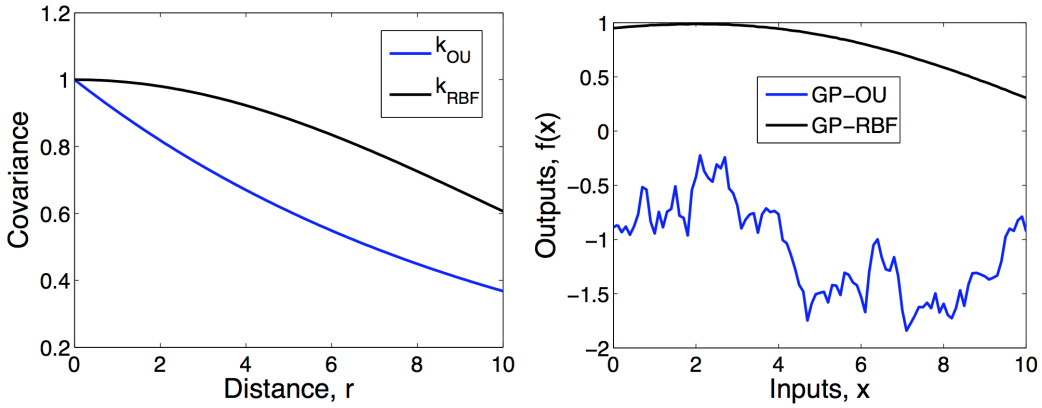
In stationary kernels, we always have a term $\tau = \mathbf{x} - \mathbf{x}'$ and the kernel is a function of τ, i.e. $k(\mathbf{x}, \mathbf{x}') = k(\tau)$. What if we want to treat $\mathbf{x} - \mathbf{x}'$ differently across the input domain? We can instead using an arbitrary warping function $g(\mathbf{x})$ so that the kernel becomes $k\left(g(\mathbf{x}), g(\mathbf{x}')\right)$. Thus we can choose the warping function as needed.

2.7 Ornstein-Uhlenbeck kernel

The RBF kernel is sometimes considered to be too smooth. We can simply replace the quadratic Euclidean distance with an absolute distance:

$$k_{OU}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{l}\right)$$

This is the Ornstein-Uhlenbeck (OU) kernel.



The absolute distance has a relatively sharp peak at zero distance (a), and the covariance decrease exponentially for a increasing distance. That makes the output no longer smooth (b).

2.8 Matern Kernel

More generally, if take the spectral densities the stationary kernel $k(\mathbf{x}, \mathbf{x}') = k(\tau)$, i.e. take the Fourier transform of RBF kernel, we have

$$S(s) = \int k(\tau) e^{-2\pi i s^{\mathrm{T}} \tau} d\tau.$$

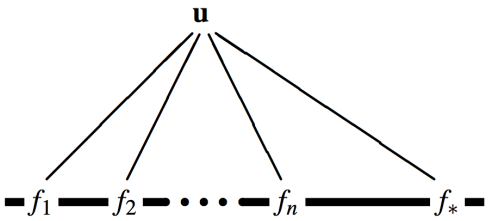
Then we apply a Student-t spectral density for $S(s)$, and take the inverse Fourier transform, we recover the Matern kernel

$$k_{Matern}(\mathbf{x}, \mathbf{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,|\mathbf{x} - \mathbf{x}'|}{l} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}\,|\mathbf{x} - \mathbf{x}'|}{l} \right),$$

where K_ν is a modified Bessel function.

By setting $\nu = 1$, we obtain the OU kernel. The Matern kernel dose not have *concentration of measure* problem for high dimensional inputs to the extent of the RBF (Gaussian) kernel. It also gives rise to a Markovian process.

3 Gaussian Process Regression



In the Gaussian Process regression all the process f_n and f_* are dependent.

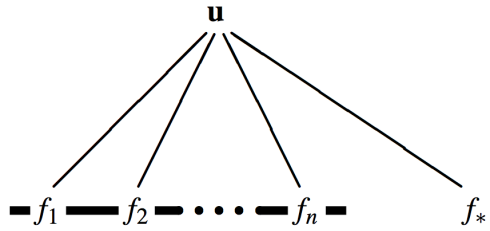
3.1 Inducing Inputs

Gaussian process $\mathbf{f}$ and $\mathbf{f}_*$ evaluated at n training points and J testing points. Assuming we introduce m ($m \ll n$) latent variables $\mathbf{u}$. The joint probability can be written as

$$p(\mathbf{f}, \mathbf{f}_*) = \int p(\mathbf{f}, \mathbf{f}_*, \mathbf{u}) d\mathbf{u} = \int p(\mathbf{f}, \mathbf{f}_* | \mathbf{u}) p(\mathbf{u}) d\mathbf{u}, \qquad \text{where} \quad \mathrm{p}(\mathbf{u}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathbf{u},\mathbf{u}}).$$

Assuming $\mathbf{f}$ and $\mathbf{f}_*$ are conditionally independent given $\mathbf{u}$:

$$p(\mathbf{f}, \mathbf{f}_*) \approx q(\mathbf{f}, \mathbf{f}_*) = \int q(\mathbf{f}|\mathbf{u}) q(\mathbf{f}_*|\mathbf{u}) p(\mathbf{u}) d\mathbf{u}.$$



So we have the exact conditional distributions:

$$p(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{K_{f,u}}\mathbf{K_{u,u}^{-1}}\mathbf{u}, \mathbf{K_{f,f}} - \mathbf{Q_{f,f}})$$

$$p(\mathbf{f_*}|\mathbf{u}) = \mathcal{N}(\mathbf{K_{f_*,u}}\mathbf{K_{u,u}^{-1}}\mathbf{u}, \mathbf{K_{f_*,f_*}} - \mathbf{Q_{f_*,f_*}})$$

$$\text{where} \quad \mathbf{Q_{a,b}} = \mathbf{K_{a,u}}\mathbf{K_{u,u}^{-1}}\mathbf{K_{u,b}}$$

3.2 Subsets of Regressors (SoR) Approximation

For any input $\mathbf{x}_*$, the corresponding function value f_* is given by:

$$f_* = \mathbf{K}_{*,\mathbf{u}}\mathbf{w_u}, \qquad \text{with} \quad \mathrm{p}(\mathbf{w_u}) = \mathcal{N}(\mathbf{0}, \mathbf{K_{u,u}^{-1}}).$$

Particularly, for $\mathbf{u}$, we have $\mathbf{u} = \mathbf{K_{u,u}}\mathbf{w_u}$ and we can redefine the SoR model in an equivalent, more intuitive way:

$$\mathbf{f}_* = \mathbf{K}_{*,\mathbf{u}}\mathbf{K}_{\mathbf{u},\mathbf{u}}^{-1}\mathbf{u}, \qquad \text{with} \quad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{K_{u,u}}).$$

Given that there is a deterministic relation between any f? and u, the approximate conditional distributions in the integral are given by:

$$q_{SoR}(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{K_{f,u}}\mathbf{K_{u,u}^{-1}}\mathbf{u}, \mathbf{0})$$

$$q_{SoR}(\mathbf{f_*}|\mathbf{u}) = \mathcal{N}(\mathbf{K}_{*,\mathbf{u}}\mathbf{K_{u,u}^{-1}}\mathbf{u}, \mathbf{0}),$$

with zero conditional covariance. Then we obtain the SoR approximation:

$$q_{SoR}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Q_{f,f}} & \mathbf{Q}_{\mathbf{f},*} \\ \mathbf{Q}_{\mathbf{f},*} & \mathbf{Q}_{*,*} \end{bmatrix}\right), \qquad \text{where} \quad \mathbf{Q_{a,b}} = \mathbf{K_{a,u}}\mathbf{K_{u,u}^{-1}}\mathbf{K_{u,b}}$$

We can finally derive the predictive conditional as:

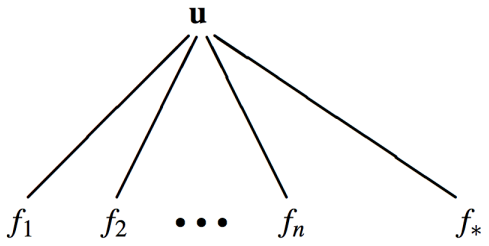
$$q_{SoR}(\mathbf{f}_*, \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{A}),$$

$$\text{where} \quad \boldsymbol{\mu} = \mathbf{Q}_{*,\mathbf{f}}(\mathbf{Q_{f,f}} + \sigma^2\mathbf{I})^{-1}\mathbf{y},$$

$$\mathbf{A} = \mathbf{Q}_{*,*} - \mathbf{Q}_{*,\mathbf{f}}(\mathbf{Q_{f,f}} + \sigma^2\mathbf{I})^{-1}\mathbf{Q}_{\mathbf{f},*}.$$

3.3 The Fully Independent Training Conditional (FITC) Approximation

Instead of ignoring the variance, FITC proposes an approximation to the training conditional distribution of f given $\mathbf{u}$ as a further independence assumption.



$$q_{FITC}(\mathbf{f}, \mathbf{u}) = \prod_{i=1}^{n} p(f_i|\mathbf{u}) = \mathcal{N}(\mathbf{K_{f,u}}\mathbf{K_{u,u}^{-1}}\mathbf{u}, \mathrm{diag}[\mathbf{K_{f,f}} - \mathbf{Q_{f,f}}])$$

$$q_{FITC}(\mathbf{f}_*, \mathbf{u}) = p(f_*|\mathbf{u})$$

Integrating away $\mathbf{u}$, The effective prior implied by the FITC is given by

$$q_{SoR}(\mathbf{f}, \mathbf{f}_*) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{Q}_{\mathbf{f},\mathbf{f}} - \text{diag}[\mathbf{K}_{\mathbf{f},\mathbf{f}} - \mathbf{Q}_{\mathbf{f},\mathbf{f}}] & \mathbf{Q}_{\mathbf{f},*} \\ \mathbf{Q}_{\mathbf{f},*} & \mathbf{K}_{*,*} \end{bmatrix}\right)$$