

CS485/685

Lecture 14: Feb 25, 2016

Gaussian Processes

[B] Section 6.4 [M] Chap. 15 [HTF]
Sec. 8.3

Gaussian Process Regression

- Idea: distribution over functions

Bayesian Linear Regression

- Setting: $f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$ and $y = f(\mathbf{x}) + \epsilon$

$\downarrow$
unknown

$\downarrow$
 $N(0, \sigma^2)$

- Weight space view:
 - Prior: $\Pr(\mathbf{w})$
 - Posterior: $\Pr(\mathbf{w}|\mathbf{X}, \mathbf{y}) = k \Pr(\mathbf{w}) \Pr(\mathbf{y}|\mathbf{w}, \mathbf{X})$

$\downarrow$
Gaussian

$\downarrow$
Gaussian

$\downarrow$
Gaussian

Bayesian Linear Regression

- Setting: $f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$ and $y = f(\mathbf{x}) + \epsilon$

$\downarrow$
 unknown

$\downarrow$
 $N(0, \sigma^2)$

- Function space view:

– Prior: $\Pr(f(\mathbf{x}_*)) = \int_{\mathbf{w}} \Pr(f|\mathbf{w}, \mathbf{x}_*) \Pr(\mathbf{w}) d\mathbf{w}$

$\downarrow$
 Gaussian

$\downarrow$
 Deterministic

$\downarrow$
 Gaussian

– Posterior: $\Pr(f(\mathbf{x}_*)|\mathbf{X}, \mathbf{y}) = \int_{\mathbf{w}} \Pr(f|\mathbf{w}, \mathbf{x}_*) \Pr(\mathbf{w}|\mathbf{X}, \mathbf{y}) d\mathbf{w}$

$\downarrow$
 Gaussian

$\downarrow$
 Deterministic

$\downarrow$
 Gaussian

Gaussian Process

- According to the function view, there is a Gaussian at $f(\mathbf{x}_*)$ for every $\mathbf{x}_*$. Those Gaussians are correlated through w .
- What is the general form of $\Pr(f)$ (i.e., distribution over functions)?
- Answer: **Gaussian Process** (infinite dimensional Gaussian distribution)

Gaussian Process

- Distribution over functions:

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \forall \mathbf{x}, \mathbf{x}'$$

- Where $m(\mathbf{x}) = E(f(\mathbf{x}))$ is the mean
and $k(\mathbf{x}, \mathbf{x}') = E((f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}')))$ is
the kernel covariance function

Mean function $m(\mathbf{x})$

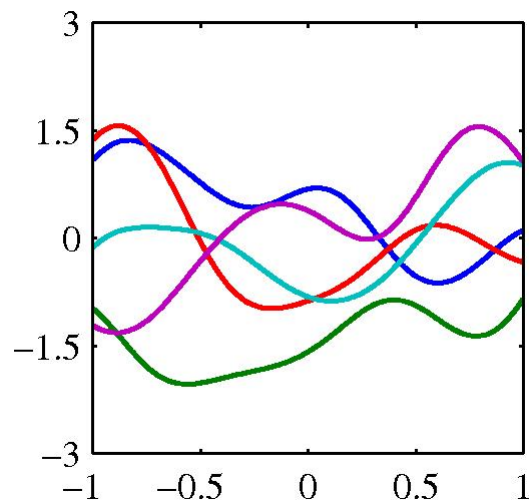
- Compute the mean function $m(\mathbf{x})$ as follows:
- Let $f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}$
with $\mathbf{w} \sim N(\mathbf{0}, \alpha^{-1} \mathbf{I})$
- Then
$$\begin{aligned} m(\mathbf{x}) &= E(f(\mathbf{x})) \\ &= E(\mathbf{w})^T \phi(\mathbf{x}) \\ &= \mathbf{0} \end{aligned}$$

Kernel covariance function $k(\mathbf{x}, \mathbf{x}')$

- Compute kernel covariance $k(\mathbf{x}, \mathbf{x}')$ as follows:
- $$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= E(f(\mathbf{x})f(\mathbf{x}')) \\ &= \phi(\mathbf{x})^T E(\mathbf{w}\mathbf{w}^T)\phi(\mathbf{x}') \\ &= \phi(\mathbf{x})^T \frac{I}{\alpha} \phi(\mathbf{x}') \\ &= \frac{\phi(\mathbf{x})^T \phi(\mathbf{x}')}{\alpha} \end{aligned}$$
- In some cases we can use domain knowledge to specify k directly.

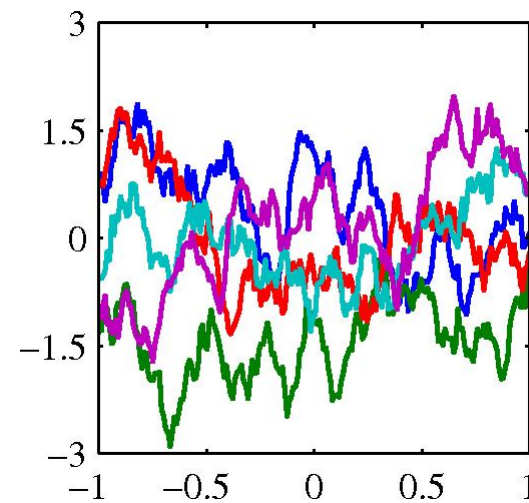
Examples

- Sampled functions from a Gaussian Process



Gaussian kernel

$$k(\mathbf{x}, \mathbf{x}') = e^{-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}}$$



Exponential kernel

(Brownian motion)

$$k(\mathbf{x}, \mathbf{x}') = e^{-\theta|\mathbf{x} - \mathbf{x}'|}$$

Gaussian Process Regression

- Gaussian Process Regression corresponds to kernelized Bayesian Linear Regression
- Bayesian Linear Regression:
 - Weight space view
 - Goal: $\Pr(\mathbf{w}|\mathbf{X}, \mathbf{y})$ (posterior over $\mathbf{w}$)
 - Complexity: cubic in # of basis functions
- Gaussian Process Regression:
 - Function space view
 - Goal: $\Pr(f|\mathbf{X}, \mathbf{y})$ (posterior over f)
 - Complexity: cubic in # of training points

Recap: Bayesian Linear Regression

- Prior: $\Pr(\mathbf{w}) = N(\mathbf{0}, \Sigma)$
- Likelihood: $\Pr(\mathbf{y}|\mathbf{X}, \mathbf{w}) = N(\mathbf{w}^T \Phi, \sigma^2 I)$
- Posterior: $\Pr(\mathbf{w}|\mathbf{X}, \mathbf{y}) = N(\bar{\mathbf{w}}, \mathbf{A}^{-1})$
where $\bar{\mathbf{w}} = \sigma^{-2} \mathbf{A}^{-1} \Phi \mathbf{y}$ and $\mathbf{A} = \sigma^{-2} \Phi \Phi^T + \Sigma^{-1}$
- Prediction:
 $\Pr(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = N(\phi(\mathbf{x}_*)^T \mathbf{A}^{-1} \Phi \mathbf{y}, \phi(\mathbf{x}_*)^T \mathbf{A}^{-1} \phi(\mathbf{x}_*))$
- Complexity: inversion of $\mathbf{A}$ is cubic in # of basis functions

Gaussian Process Regression

- Prior: $\Pr(f(\cdot)) = N(m(\cdot), k(\cdot, \cdot))$
- Likelihood: $\Pr(\mathbf{y}|\mathbf{X}, f) = N(f(\cdot), \sigma^2 \mathbf{I})$
- Posterior: $\Pr(f(\cdot)|\mathbf{X}, \mathbf{y}) = N(\bar{f}(\cdot), k'(\cdot, \cdot))$
where $\bar{f}(\cdot) = k(\cdot, \mathbf{X})(\mathbf{K} + \sigma^{-2} \mathbf{I})^{-1} \mathbf{y}$ and
$$k'(\cdot, \cdot) = k(\cdot, \cdot) - k(\cdot, \mathbf{X})(\mathbf{K} + \sigma^2 \mathbf{I})^{-1} k(\mathbf{X}, \cdot)$$
- Prediction: $\Pr(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = N(\bar{f}(\mathbf{x}_*), k'(\mathbf{x}_*, \mathbf{x}_*))$
- Complexity: inversion of $\mathbf{K} + \sigma^2 \mathbf{I}$ is cubic in # of training points

Case Study: AIBO Gait Optimization



Gait Optimization

- Problem: find best parameter setting of the gait controller to maximize walking speed
 - Why?: Fast robots have a better chance of winning in robotic soccer
- Solutions:
 - Stochastic hill climbing
 - **Gaussian Processes**
 - Lizotte, Wang, Bowling, Schuurmans (2007) Automatic Gait Optimization with Gaussian Processes, *International Joint Conferences on Artificial Intelligence (IJCAI)*.

Search Problem

- Let $\mathbf{x} \in \mathbb{R}^{15}$, be a vector of 15 parameters that defines a controller for gait
- Let $f: \mathbf{x} \rightarrow \mathbb{R}$ be a mapping from controller parameters to gait speed
- Problem: find parameters $\mathbf{x}^*$ that yield highest speed.

$$\mathbf{x}^* \leftarrow \operatorname{argmax}_{\mathbf{x}} f(\mathbf{x})$$

But f is unknown...

Approach

- Picture

Approach

- Initialize $f(\cdot) \sim GP(m(\cdot), k(\cdot, \cdot))$
- Repeat:

- Select new $\mathbf{x}$:

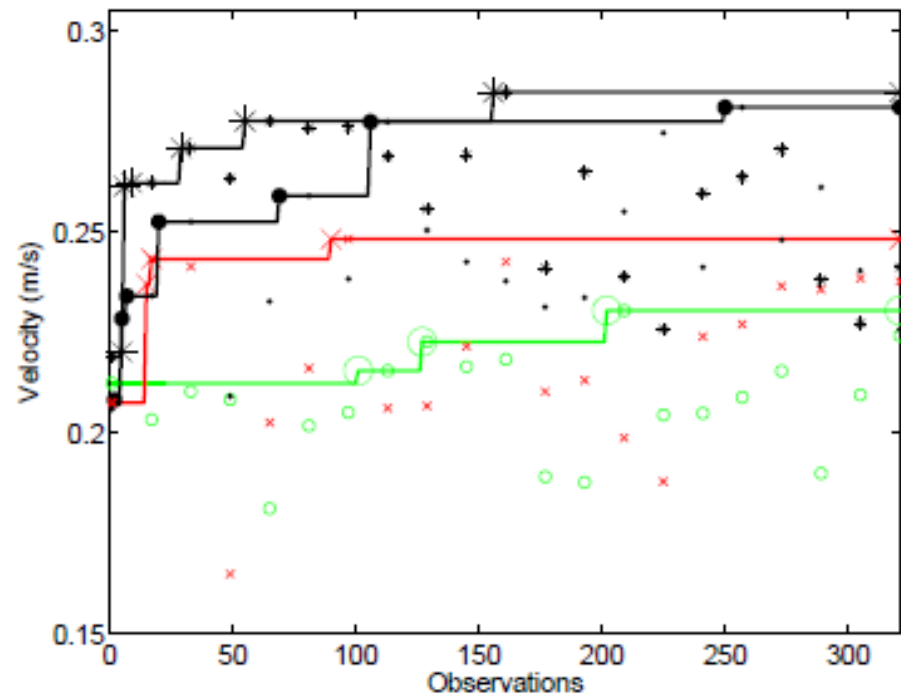
$$\mathbf{x}_{new} \leftarrow \underset{\mathbf{x}' \in X}{\operatorname{argmax}} \frac{k(\mathbf{x}, \mathbf{x})}{f(\mathbf{x}') - m(\mathbf{x})}$$

- Evaluate $f(\mathbf{x}_{new})$ by observing speed of robot with parameters set to $\mathbf{x}_{new}$

- Update Gaussian process:

- $\mathbf{X} \leftarrow \mathbf{X} \cup \{\mathbf{x}_{new}\}$ and $\mathbf{y} \leftarrow \mathbf{y} \cup f(\mathbf{x}_{new})$
 - $m(\cdot) \leftarrow k(\cdot, \mathbf{X})(\mathbf{K} + \sigma^{-2}\mathbf{I})^{-1}\mathbf{y}$
 - $k(\cdot, \cdot) \leftarrow k(\cdot, \cdot) - k(\cdot, \mathbf{X})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}k(\mathbf{X}, \cdot)$

Results



(●) GP w/MPI	(*) GP w/MPI	(×) H.Clmb	(○) U.Rand
0.281 m/s	0.285 m/s	0.248 m/s	0.230 m/s
$\sigma_f^2 = 0.06$	$\sigma_f^2 = 0.6$		

Gaussian kernel:

$$k(x, x') = \sigma_f^2 e^{-\frac{1}{2}(x-x')^T S(x-x')}$$