Modern Gaussian Processes: Scalable Inference and Novel Applications

(Part I-b) Introduction to Gaussian Processes

Edwin V. Bonilla and **Maurizio Filippone**

CSIRO's Data61, Sydney, Australia and EURECOM, Sophia Antipolis, France

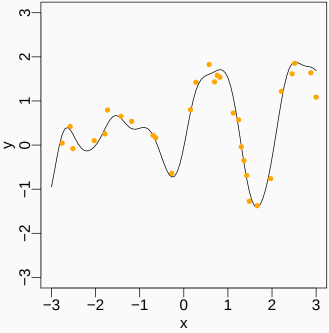
July 14th, 2019

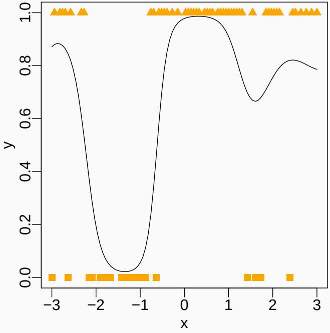
Outline

Bayesian Modeling

Learning from Data — Function Estimation

- Take these two examples

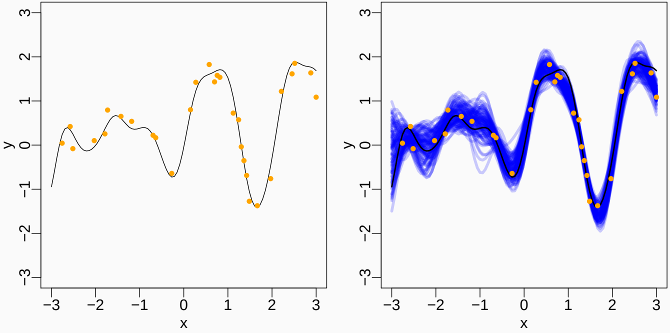




- We are interested in estimating a function $\mathbf{f}(\mathbf{x})$ from data
- Most problems in Machine Learning can be cast this way!

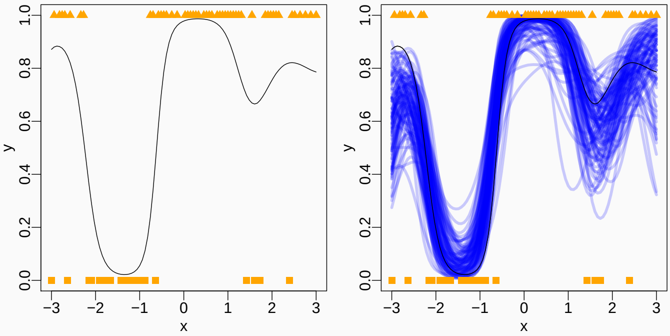
What do Bayesian Models Have to Offer?

- Regression example



What do Bayesian Models Have to Offer?

- Classification example



Gaussian Processes

Linear Models

- Implement a linear combination of basis functions

$$\mathbf{f}(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\varphi}(\mathbf{x})$$

with

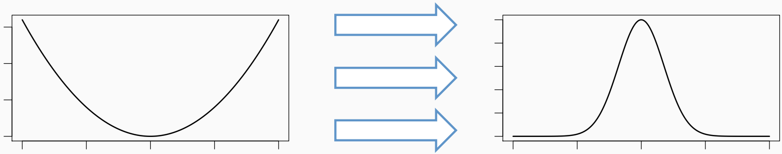
$$\boldsymbol{\varphi}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_D(\mathbf{x}))^\top$$

Probabilistic Interpretation of Loss Minimization

- Inputs : $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top$
- Labels : $\mathbf{y} = (y_1, \ldots, y_N)^\top$
- Weights : $\mathbf{w} = (w_1, \ldots, w_D)^\top$

Quadratic Loss

$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) \propto \exp(-\mathrm{Loss})$

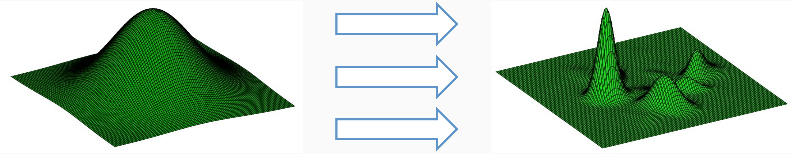


- Minimization of a loss function
- ... equivalent as maximizing likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$

Bayesian Inference

- Inputs : $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top$
- Labels : $\mathbf{y} = (y_1, \ldots, y_N)^\top$
- Weights : $\mathbf{w} = (w_1, \ldots, w_D)^\top$

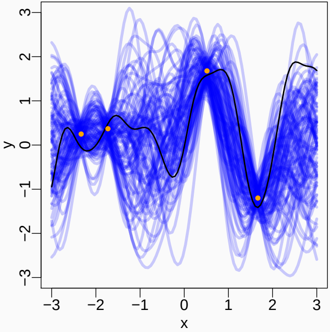
$p(\mathbf{w})$ $\qquad\qquad$ $p(\mathbf{w}|\mathbf{y}, \mathbf{X})$



$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}}$$

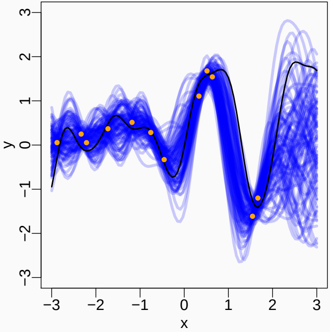
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



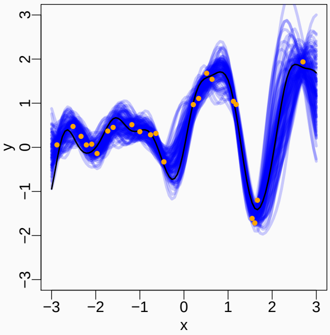
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



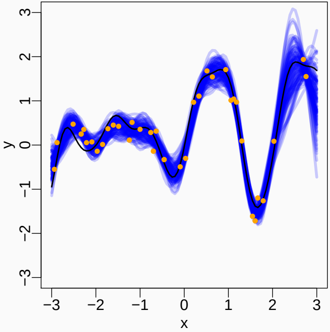
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



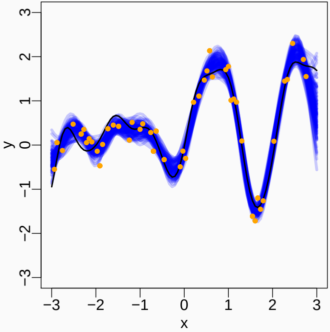
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



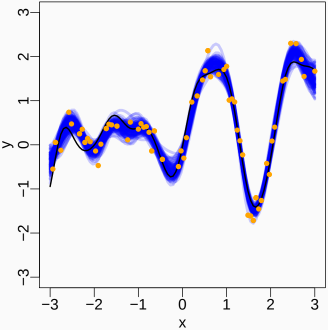
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



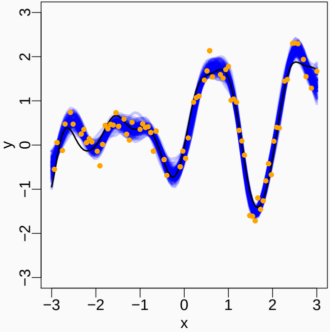
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



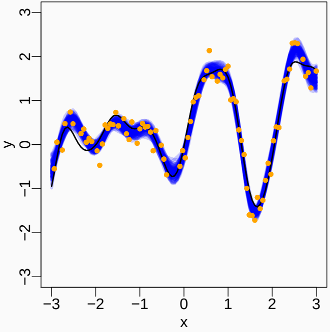
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



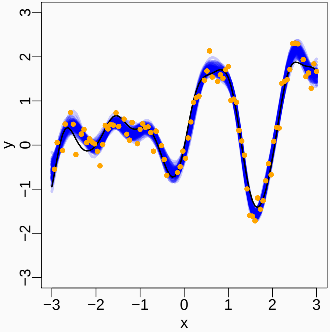
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



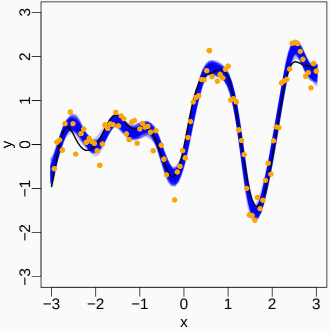
Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



Bayesian Linear Models in Action

- Today's posterior is tomorrow's prior



Bayesian Linear Regression

- Modeling observations as noisy realizations of a linear combination of the features:

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{\Phi w}, \sigma^2 \mathbf{I})$$

- $\mathbf{\Phi} = \mathbf{\Phi}(\mathbf{X})$ has entries

$$\mathbf{\Phi} = \begin{bmatrix} \varphi_1(\mathbf{x}_1) & \dots & \varphi_D(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{x}_N) & \dots & \varphi_D(\mathbf{x}_N) \end{bmatrix}$$

Bayesian Linear Regression

- Modeling observations as noisy realizations of a linear combination of the features:

$$p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = \mathcal{N}(\mathbf{\Phi}\mathbf{w}, \sigma^2\mathbf{I})$$

- $\mathbf{\Phi} = \mathbf{\Phi}(\mathbf{X})$ has entries

$$\mathbf{\Phi} = \begin{bmatrix} \varphi_1(\mathbf{x}_1) & \dots & \varphi_D(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{x}_N) & \dots & \varphi_D(\mathbf{x}_N) \end{bmatrix}$$

- Gaussian prior over model parameters:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \mathbf{S})$$

Bayesian Linear Regression

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X},\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{y}|\mathbf{X},\mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

Bayesian Linear Regression

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
  - Distribution over parameters *after* observing data

Bayesian Linear Regression

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
  - Distribution over parameters *after* observing data
- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of "fitness"

Bayesian Linear Regression

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
  - Distribution over parameters *after* observing data
- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of "fitness"
- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data

Bayesian Linear Regression

- Bayes rule:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{\int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})d\mathbf{w}} = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}$$

- **Posterior density**: $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$
  - Distribution over parameters *after* observing data
- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$
  - Measure of "fitness"
- **Prior density**: $p(\mathbf{w})$
  - Anything we know about parameters *before* we see any data
- **Marginal likelihood**: $p(\mathbf{y}|\mathbf{X})$
  - It is a normalization constant – ensures $\int p(\mathbf{w}|\mathbf{X}, \mathbf{y})\ d\mathbf{w} = 1$.

Bayesian Linear Regression

- Posterior **must be** Gaussian (Proof in the Appendix)

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Covariance:

$$\boldsymbol{\Sigma} = \left(\frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \mathbf{S}^{-1}\right)^{-1}$$

- Mean:

$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^\top\mathbf{y}$$

- Predictions – with a similar tedious exercise...

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(\mathbf{x}_*^\top\boldsymbol{\mu}, \sigma^2 + \mathbf{x}_*^\top\boldsymbol{\Sigma}\mathbf{x}_*)$$

Gaussian Processes

- Linear models require specifying a set of basis functions
  - Polynomials, Trigonometric, ...??

Gaussian Processes

- Linear models require specifying a set of basis functions
  - Polynomials, Trigonometric, ...??
- Gaussian Processes work implicitly with a possibly infinite set of basis functions!

Bayesian Linear Regression as a Kernel Machine

- Predictions can be expressed exclusively in terms of scalar products as follows

$$k(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\psi}(\mathbf{x}_i)^\top \boldsymbol{\psi}(\mathbf{x}_j)$$

- This allows us to work with either $k(\cdot, \cdot)$ or $\boldsymbol{\psi}(\cdot)$
- Why is this useful??

Bayesian Linear Regression as a Kernel Machine

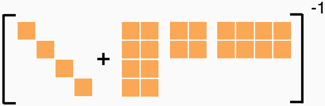
- Working with $\psi(\cdot)$ costs $O(D^2)$ storage, $O(D^3)$ time
- Working with $k(\cdot, \cdot)$ costs $O(N^2)$ storage, $O(N^3)$ time

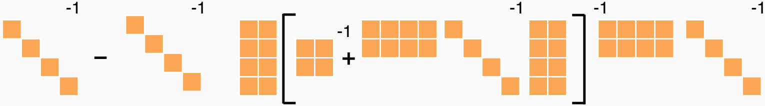
Bayesian Linear Regression as a Kernel Machine

Proof sketch - more in the Appendix

- To show that Bayesian Linear Regression can be formulated through scalar products only, we need Woodbury identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$





Bayesian Linear Regression as a Kernel Machine

Proof sketch - more in the Appendix

- Woodbury identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

- We can rewrite:

$$\begin{aligned}
\mathbf{\Sigma} &= \left(\frac{1}{\sigma^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{S}^{-1}\right)^{-1} \\
&= \mathbf{S} - \mathbf{S}\mathbf{\Phi}^\top\left(\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{S}\mathbf{\Phi}^\top\right)^{-1}\mathbf{\Phi}\mathbf{S}
\end{aligned}$$

- We set $A = \mathbf{S}$, $U = V^\top = \mathbf{\Phi}^\top$, and $C = \frac{1}{\sigma^2}\mathbf{I}$

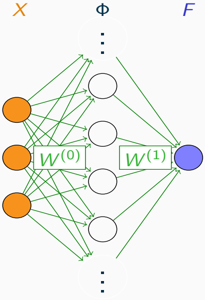
Kernels

- We can pick $k(\cdot, \cdot)$ so that $\psi(\cdot)$ is infinite dimensional!
- It is possible to show that for

  $$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2}\right)$$

  there exists a corresponding $\psi(\cdot)$ that is infinite dimensional!
  (Proof in the Appendix)
- There are other kernels satisfying this property

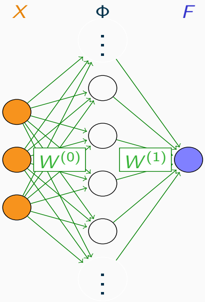
Gaussian Processes as Infinitely-Wide Shallow Neural Nets



- Take $W^{(i)} \sim \mathcal{N}(\mathbf{0}, \alpha_i I)$
- Central Limit Theorem implies that $\mathbf{f}$ is Gaussian
- $\mathbf{f}$ has zero-mean
- $\mathrm{cov}(\mathbf{f}) = \mathbb{E}_{p(W^{(0)}, W^{(1)})}[\boldsymbol{\Phi}(\mathbf{X}W^{(0)})W^{(1)}W^{(1)\top}\boldsymbol{\Phi}(\mathbf{X}W^{(0)})^\top]$

Neal, *LNS*, 1996

Gaussian Processes as Infinitely-Wide Shallow Neural Nets



- Take $W^{(i)} \sim \mathcal{N}(\mathbf{0}, \alpha_i I)$
- Central Limit Theorem implies that $\mathbf{f}$ is Gaussian

- $\mathbf{f}$ has zero-mean
- $\mathrm{cov}(\mathbf{f}) = \alpha_1 \mathbb{E}_{p(W^{(0)})}[\mathbf{\Phi}(\mathbf{X}W^{(0)})\mathbf{\Phi}(\mathbf{X}W^{(0)})^\top]$
- Some choices of $\mathbf{\Phi}$ lead to analytic expression of known kernels (RBF, Matérn, arc-cosine, Brownian motion, ...)

Neal, *LNS*, 1996

- Latent function:

$$\mathbf{f} = \mathbf{w}^\top \varphi(\mathbf{x})$$

  with $\varphi(\cdot)$ possibly infinite dimensional!
- The choice of $\varphi(\cdot)$ and the prior over $\mathbf{w}$ induce a distribution over functions

Definition

f is distributed according to a Gaussian process iff for any subset $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ the evaluation of f is jointly Gaussian

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')) \qquad \text{then}$$

$$\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$$

Gaussian Processes for Regression

- Bayes rule:

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

Gaussian Processes for Regression

- Bayes rule:

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

- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I})$

Gaussian Processes for Regression

- Bayes rule:

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

- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{0},\sigma^2\mathbf{I})$
- **Prior over latent variables**: Implied by the prior over $\mathbf{w}$

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

Gaussian Processes for Regression

- Bayes rule:

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

- **Conditional Likelihood** : $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I})$
- **Prior over latent variables**: Implied by the prior over $\mathbf{w}$

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

- **Marginal likelihood**: $p(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I})$

Optimization of Gaussian Process parameters

- The kernel has parameters that have to be tuned

$$k(\mathbf{x}_i, \mathbf{x}_j) = \alpha \exp(-\beta \|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

  ... and there is also the noise parameter σ^2.

- Define $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)$
- How should we tune them?

Optimization of Gaussian Process parameters

- Define $\mathbf{K_y} = \mathbf{K} + \sigma^2 \mathbf{I}$
- Maximize the logarithm of the marginal likelihood

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \mathbf{K_y})$$

that is

$$-\frac{1}{2} \log |\mathbf{K_y}| - \frac{1}{2} \mathbf{y}^\top \mathbf{K_y}^{-1} \mathbf{y} + \text{const.}$$

- Derivatives can be useful for gradient-based optimization

$$\frac{\partial \log[p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}_i}$$

Optimization of Gaussian Process parameters

- Log-marginal likelihood

$$-\frac{1}{2}\log|\mathbf{K_y}| - \frac{1}{2}\mathbf{y}^\top \mathbf{K_y}^{-1}\mathbf{y} + \text{const.}$$

- Derivatives can be useful for gradient-based optimization:

$$\frac{\partial \log[p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}_i} = -\frac{1}{2}\mathrm{Tr}\left(\mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\right) + \frac{1}{2}\mathbf{y}^\top \mathbf{K_y}^{-1}\frac{\partial \mathbf{K_y}}{\partial \boldsymbol{\theta}_i}\mathbf{K_y}^{-1}\mathbf{y}$$

Challenges

Challenges

- Non-Gaussian Likelihoods?
- Scalability?
- Kernel design?

Marginal likelihood of GP models - non-Gaussian case

- Marginal likelihood

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

  can only be computed if $p(\mathbf{y}|\mathbf{f})$ is Gaussian
- What if $p(\mathbf{y}|\mathbf{f})$ is **not** Gaussian?

Scalability

- Marginal likelihood

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

  can only be computed if $p(\mathbf{y}|\mathbf{X},\mathbf{f})$ is Gaussian

- ... even then

$$\log[p(\mathbf{y}|\mathbf{X},\theta)] = -\frac{1}{2}\log|\mathbf{K_y}| - \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K_y}^{-1}\mathbf{y} + \mathrm{const.}$$

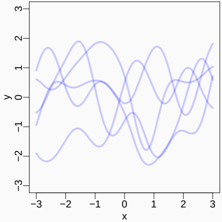
  where $\mathbf{K_y} = \mathbf{K}(\mathbf{X},\theta)$ is a $N \times N$ dense matrix!

- Complexity of exact method is $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ space!

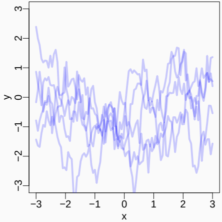
Kernel Design

- The choice of a kernel is critical for good performance
- This encodes any assumptions on the prior over functions

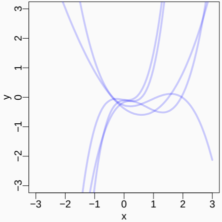
RBF



Matérn



Polynomial



Appendix

Kernels

- For simplicity consider one dimensional inputs x_i, x_j
- Expand the Gaussian kernel $k(x_i, x_j)$ as

$$\exp\left(-\frac{(x_i - x_j)^2}{2}\right) = \exp\left(-\frac{x_i^2}{2}\right)\exp\left(-\frac{x_j^2}{2}\right)\exp\left(x_i x_j\right)$$

- Focusing on the last term and applying the Taylor expansion of the $\exp(\cdot)$ function

$$\exp\left(x_i x_j\right) = 1 + \left(x_i x_j\right) + \frac{\left(x_i x_j\right)^2}{2!} + \frac{\left(x_i x_j\right)^3}{3!} + \frac{\left(x_i x_j\right)^4}{4!} + \dots$$

Kernels

- Define the infinite dimensional mapping

$$\boldsymbol{\psi}(x) = \exp\left(-\frac{x^2}{2}\right)\left(1, x, \frac{x^2}{\sqrt{2!}}, \frac{x^3}{\sqrt{3!}}, \frac{x^4}{\sqrt{4!}}, \dots\right)^\top$$

- It is easy to verify that

$$k(x_i, x_j) = \exp\left(-\frac{(x_i - x_j)^2}{2}\right) = \boldsymbol{\psi}(x_i)^\top \boldsymbol{\psi}(x_j)$$

Bayesian Linear Regression - Finding posterior parameters

- Ignoring normalizing constants, the posterior is:

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) &\propto \exp\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right\} \\
&= \exp\left\{-\frac{1}{2}(\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\}
\end{aligned}
$$

Bayesian Linear Regression - Finding posterior parameters

- Ignoring non-$\mathbf{w}$ terms, the prior multiplied by the likelihood is:

$$
\begin{aligned}
& p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) \\
\propto \quad & \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w})^\top(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w})\right\} \exp\left\{-\frac{1}{2}\mathbf{w}^\top \mathbf{S}^{-1}\mathbf{w}\right\} \\
\propto \quad & \exp\left\{-\frac{1}{2}\left(\mathbf{w}^\top \left[\frac{1}{\sigma^2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \mathbf{S}^{-1}\right]\mathbf{w} - \frac{2}{\sigma^2}\mathbf{w}^\top\boldsymbol{\Phi}^\top\mathbf{y}\right)\right\}
\end{aligned}
$$

- Posterior (from previous slide):

$$\propto \exp\left\{-\frac{1}{2}(\mathbf{w}^\top\boldsymbol{\Sigma}^{-1}\mathbf{w} - 2\mathbf{w}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\right\}$$

Bayesian Linear Regression - Finding posterior parameters

- Equate individual terms on each side.
- Covariance:

$$\begin{aligned} \mathbf{w}^\top \mathbf{\Sigma}^{-1} \mathbf{w} &= \mathbf{w}^\top \left[\frac{1}{\sigma^2} \mathbf{\Phi}^\top \mathbf{\Phi} + \mathbf{S}^{-1} \right] \mathbf{w} \\ \mathbf{\Sigma} &= \left(\frac{1}{\sigma^2} \mathbf{\Phi}^\top \mathbf{\Phi} + \mathbf{S}^{-1} \right)^{-1} \end{aligned}$$

- Mean:

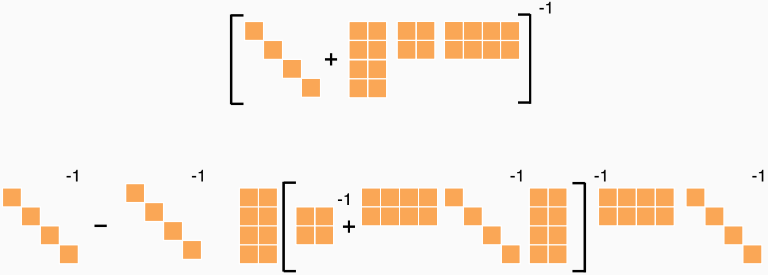
$$\begin{aligned} 2\mathbf{w}^\top \mathbf{\Sigma}^{-1} \boldsymbol{\mu} &= \frac{2}{\sigma^2} \mathbf{w}^\top \mathbf{\Phi}^\top \mathbf{y} \\ \boldsymbol{\mu} &= \frac{1}{\sigma^2} \mathbf{\Sigma} \mathbf{\Phi}^\top \mathbf{y} \end{aligned}$$

Bayesian Linear Regression as a Kernel Machine

- To show that Bayesian Linear Regression can be formulated through scalar products only, we need Woodbury identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

- Intuitively:



Bayesian Linear Regression as a Kernel Machine

- Woodbury identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

- We can rewrite:

$$\begin{aligned}\mathbf{\Sigma} &= \left(\frac{1}{\sigma^2}\mathbf{\Phi}^\top\mathbf{\Phi} + \mathbf{S}^{-1}\right)^{-1} \\ &= \mathbf{S} - \mathbf{S}\mathbf{\Phi}^\top\left(\sigma^2\mathbf{I} + \mathbf{\Phi}\mathbf{S}\mathbf{\Phi}^\top\right)^{-1}\mathbf{\Phi}\mathbf{S}\end{aligned}$$

- We set $A = \mathbf{S}$, $U = V^\top = \mathbf{\Phi}^\top$, and $C = \frac{1}{\sigma^2}\mathbf{I}$

Bayesian Linear Regression as a Kernel Machine

- Mean and variance of the predictions:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(\boldsymbol{\phi}_*^\top \boldsymbol{\mu}, \sigma^2 + \boldsymbol{\phi}_*^\top \boldsymbol{\Sigma} \boldsymbol{\phi}_*)$$

- Rewrite the variance:

$$\begin{aligned} \sigma^2 \quad &+ \quad \boldsymbol{\phi}_*^\top \boldsymbol{\Sigma} \boldsymbol{\phi}_* = \\ \sigma^2 \quad &+ \quad \boldsymbol{\phi}_*^\top \mathbf{S} \boldsymbol{\phi}_* - \boldsymbol{\phi}_*^\top \mathbf{S} \boldsymbol{\Phi}^\top \left(\sigma^2 \mathbf{I} + \boldsymbol{\Phi} \mathbf{S} \boldsymbol{\Phi}^\top\right)^{-1} \boldsymbol{\Phi} \mathbf{S} \boldsymbol{\phi}_* \end{aligned}$$

... continued

Bayesian Linear Regression as a Kernel Machine

- Mean and variance of the predictions:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(\phi_*^\top \boldsymbol{\mu}, \sigma^2 + \phi_*^\top \boldsymbol{\Sigma} \phi_*)$$

- Rewrite the variance:

$$\begin{aligned} \sigma^2 \quad &+ \quad \phi_*^\top \mathbf{S} \phi_* - \phi_*^\top \mathbf{S} \boldsymbol{\Phi}^\top \left(\sigma^2 \mathbf{I} + \boldsymbol{\Phi} \mathbf{S} \boldsymbol{\Phi}^\top\right)^{-1} \boldsymbol{\Phi} \mathbf{S} \phi_* = \\ \sigma^2 \quad &+ \quad k_{**} - \mathbf{k}_*^\top \left(\sigma^2 \mathbf{I} + \mathbf{K}\right)^{-1} \mathbf{k}_* \end{aligned}$$

- Where the mapping defining the kernel is

$$\psi(\mathbf{x}) = \mathbf{S}^{1/2} \phi(\mathbf{x})$$

and

$$\begin{aligned} k_{**} \quad &= \quad k(\mathbf{x}_*, \mathbf{x}_*) = \psi(\mathbf{x}_*)^\top \psi(\mathbf{x}_*) \\ (\mathbf{k}_*)_i \quad &= \quad k(\mathbf{x}_*, \mathbf{x}_i) = \psi(\mathbf{x}_*)^\top \psi(\mathbf{x}_i) \\ (\mathbf{K})_{ij} \quad &= \quad k(\mathbf{x}_i, \mathbf{x}_j) = \psi(\mathbf{x}_i)^\top \psi(\mathbf{x}_j) \end{aligned}$$

Bayesian Linear Regression as a Kernel Machine

- Mean and variance of the predictions:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*, \sigma^2) = \mathcal{N}(\boldsymbol{\phi}_*^\top \boldsymbol{\mu}, \sigma^2 + \boldsymbol{\phi}_*^\top \boldsymbol{\Sigma} \boldsymbol{\phi}_*)$$

- Rewrite the mean:

$$\begin{aligned}
\boldsymbol{\phi}_*^\top \boldsymbol{\mu} &= \frac{1}{\sigma^2} \boldsymbol{\phi}_*^\top \boldsymbol{\Sigma} \boldsymbol{\Phi}^\top \mathbf{y} \\
&= \frac{1}{\sigma^2} \boldsymbol{\phi}_*^\top \left(\mathbf{S} - \mathbf{S}\boldsymbol{\Phi}^\top \left(\sigma^2 \mathbf{I} + \boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\top \right)^{-1} \boldsymbol{\Phi}\mathbf{S} \right) \boldsymbol{\Phi}^\top \mathbf{y} \\
&= \frac{1}{\sigma^2} \boldsymbol{\phi}_*^\top \mathbf{S}\boldsymbol{\Phi}^\top \left(\mathbf{I} - \left(\sigma^2 \mathbf{I} + \boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\top \right)^{-1} \boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\top \right) \mathbf{y} \\
&= \frac{1}{\sigma^2} \boldsymbol{\phi}_*^\top \mathbf{S}\boldsymbol{\Phi}^\top \left(\mathbf{I} - \left(\mathbf{I} + \frac{\boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\top}{\sigma^2} \right)^{-1} \frac{\boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\top}{\sigma^2} \right) \mathbf{y}
\end{aligned}$$

... continued

Bayesian Linear Regression as a Kernel Machine

- Define $\mathbf{H} = \frac{\mathbf{\Phi S \Phi}^\top}{\sigma^2}$
- The term in the parenthesis

$$\left(\mathbf{I} - \left(\mathbf{I} + \frac{\mathbf{\Phi S \Phi}^\top}{\sigma^2}\right)^{-1} \frac{\mathbf{\Phi S \Phi}^\top}{\sigma^2}\right)$$

becomes

$$\left(\mathbf{I} - (\mathbf{I} + \mathbf{H})^{-1}\mathbf{H}\right) = \mathbf{I} - (\mathbf{H}^{-1} + \mathbf{I})^{-1}$$

- Using Woodbury (A, U, $V = \mathbf{I}$ and $C = \mathbf{H}^{-1}$)

$$\mathbf{I} - (\mathbf{H}^{-1} + \mathbf{I})^{-1} = (\mathbf{I} + \mathbf{H})^{-1}$$

Bayesian Linear Regression as a Kernel Machine

- Substituting into the expression of the predictive mean

$$
\begin{aligned}
\phi_*^\top \boldsymbol{\mu} &= \frac{1}{\sigma^2}\phi_*^\top \mathbf{S}\boldsymbol{\Phi}^\top \left(\mathbf{I} - \left(\mathbf{I} + \frac{\boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\top}{\sigma^2}\right)^{-1} \frac{\boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\top}{\sigma^2}\right)\mathbf{y} \\
&= \frac{1}{\sigma^2}\phi_*^\top \mathbf{S}\boldsymbol{\Phi}^\top \left(\mathbf{I} + \frac{\boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\top}{\sigma^2}\right)^{-1} \mathbf{y} \\
&= \phi_*^\top \mathbf{S}\boldsymbol{\Phi}^\top \left(\sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{S}\boldsymbol{\Phi}^\top\right)^{-1} \mathbf{y} \\
&= \mathbf{k}_*^\top \left(\sigma^2\mathbf{I} + \mathbf{K}\right)^{-1} \mathbf{y}
\end{aligned}
$$

- All definitions as in the case of the variance

$$
\begin{aligned}
\psi(\mathbf{X}) &= \mathbf{S}^{1/2}\phi(\mathbf{X}) \\
(\mathbf{k}_*)_i &= k(\mathbf{x}_*, \mathbf{x}_i) = \psi(\mathbf{x}_*)^\top \psi(\mathbf{x}_i) \\
(\mathbf{K})_{ij} &= k(\mathbf{x}_i, \mathbf{x}_j) = \psi(\mathbf{x}_i)^\top \psi(\mathbf{x}_j)
\end{aligned}
$$