Gaussian processes

1. Discriminative classification

2. Bayesian logistic regression & Laplace approximation

3. Generative classification

Motivating GPs from linear regression

- In linear regression, had $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$
- Gaussian prior on $\mathbf{w}$: $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$
- Joint distribution of N outputs $y_n \equiv y(\mathbf{x}_n)$?
- If $\mathbf{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$, $\Phi_{nj} = \phi_j(\mathbf{x}_n)$, then $\mathbf{y} = \mathbf{\Phi}\mathbf{w}$
- Gaussian linear model, so $\mathbf{y}$ has Gaussian distribution:

$$\mathbb{E}[\mathbf{y}] = \mathbf{0}, \qquad \mathbb{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = \mathbf{\Phi}\mathbb{E}[\mathbf{w}\mathbf{w}^{\mathrm{T}}]\mathbf{\Phi}^{\mathrm{T}} = \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$$

- $\mathbf{K} = \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ has entries $\sum_j \Phi_{nj}\Phi_{mj} = \alpha^{-1}\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_m)$
- Each entry is just as function of $\mathbf{x}_n$, $\mathbf{x}_m$

Generalization: Priors over functions

- Rephrase prior $p(\mathbf{w})$ as prior $p(y)$ over functions $y(\mathbf{x})$
- The function $y(\mathbf{x})$ is also called a stochastic process

Definition

We say $p(y)$ is a Gaussian process prior, or $y(\mathbf{x})$ is a GP under the prior, if for any N the distribution of $\mathbf{y} = (y(\mathbf{x}_1), \ldots, y(\mathbf{x}_N))^{\mathrm{T}}$ is

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{m}, \mathbf{K}) \quad \text{with} \quad K_{nm} = k(\mathbf{x}_n, \mathbf{x}_m), \quad m_n = \mu(\mathbf{x}_n)$$

- Mean function $\mu(\mathbf{x})$, mostly set to zero
- Covariance function or kernel $k(\mathbf{x}, \mathbf{x}')$
- A kernel $k(\mathbf{x}, \mathbf{x}')$ is valid if the Gram matrix $\mathbf{K}$ is positive (semi-)definite for all choices of the $\mathbf{x}_1, \ldots, \mathbf{x}_N$

Gaussian processes

1. Discriminative classification
2. Bayesian logistic regression & Laplace approximation
3. Generative classification

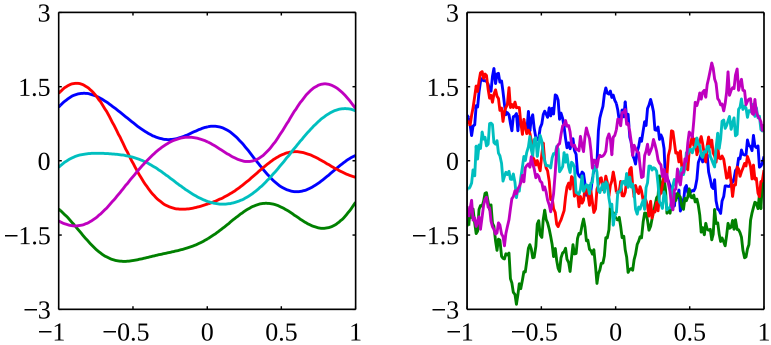
Constructing valid kernels

- So far: any $k(\mathbf{x},\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')$ is valid: scalar product of basis function vectors (also: 'feature vectors')
- If $k_1(\mathbf{x},\mathbf{x}')$ and $k_2(\mathbf{x},\mathbf{x}')$ are valid, also sum $k_1(\mathbf{x},\mathbf{x}') + k_2(\mathbf{x},\mathbf{x}')$ is
  (covariance function of $y_1(\mathbf{x}) + y_2(\mathbf{x})$)
- Similarly product $k_1(\mathbf{x},\mathbf{x}')k_2(\mathbf{x},\mathbf{x}')$
  (covariance function of $y_1(\mathbf{x})y_2(\mathbf{x})$)
- Multiplication by function of single input: $f(\mathbf{x})k_1(\mathbf{x},\mathbf{x}')f(\mathbf{x}')$
  (covariance function of $f(\mathbf{x})y_1(\mathbf{x})$)
- Multiplication by positive constant: $ck_1(\mathbf{x},\mathbf{x}')$
  (special case $f(\mathbf{x}) = \sqrt{c}$)
- Polynomial with positive coefficients $q(k_1(\mathbf{x},\mathbf{x}'))$
  (take products to get monomials, then sum)
- Exponential $\exp(k_1(\mathbf{x},\mathbf{x}'))$
  (infinite polynomial with positive coefficients)

Examples of valid kernels

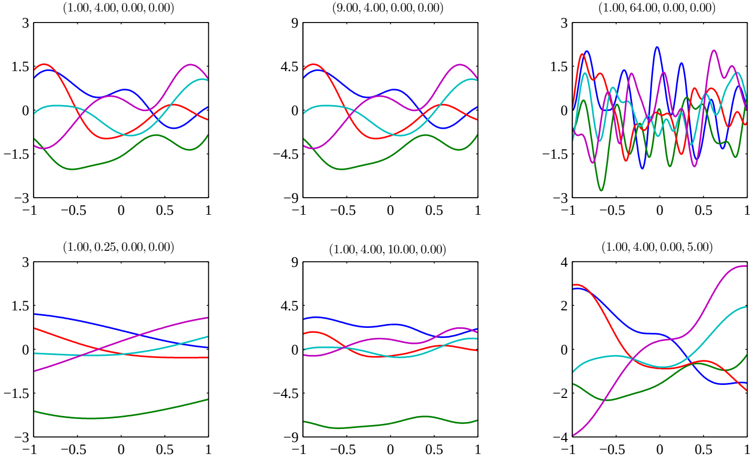
- Dot product: $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^{\mathrm{T}}\mathbf{x}'$
- Squared exponential or RBF kernel:
  $k(\mathbf{x}, \mathbf{x}') = \exp[-||\mathbf{x} - \mathbf{x}'||^2/(2\sigma^2)]$
- Ornstein-Uhlenbeck (OU) kernel:
  $k(\mathbf{x}, \mathbf{x}') = \exp(-||\mathbf{x} - \mathbf{x}'||/\sigma)$
  Superposition of infinitely many RBF kernels
  $(e^{-||\mathbf{x}-\mathbf{x}'||/\sigma} = (2/\pi)^{1/2} \int_0^\infty ds\, e^{-s^2/2} e^{-||\mathbf{x}-\mathbf{x}'||^2/(2s^2\sigma^2)})$
- Kernels from generative models: e.g. $k(\mathbf{x}, \mathbf{x}') = p(\mathbf{x})p(\mathbf{x}')$ or $k(\mathbf{x}, \mathbf{x}') = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{x}'|\mathbf{z})p(\mathbf{z})$
- Inputs $\mathbf{x}$ don't need to be vectors: strings, sets, ...

Samples from GP priors – Smoothness



Left: RBF kernel, right: OU kernel

Samples from GP priors – Effects of parameters



$$k(\mathbf{x}, \mathbf{x}') = \theta_0 \exp\left(-\tfrac{1}{2}\theta_1 ||\mathbf{x} - \mathbf{x}'||^2\right) + \theta_2 + \theta_3 \mathbf{x}^{\mathrm{T}} \mathbf{x}'$$

Gaussian processes

1. Discriminative classification
2. Bayesian logistic regression & Laplace approximation
3. Generative classification

Regression with GPs: Likelihood

- Already have prior (GP) on "clean" function $y(\mathbf{x})$
- Noise model as before: $t_n = y_n + \epsilon_n$ with $y_n = y(\mathbf{x}_n)$ and ϵ_n i.i.d. noise
- For Gaussian noise, $p(t_n|y_n) = \mathcal{N}(t_n|y_n, \beta^{-1})$
- Gives for N training outputs $\mathbf{t} = (t_1, \dots, t_N)$

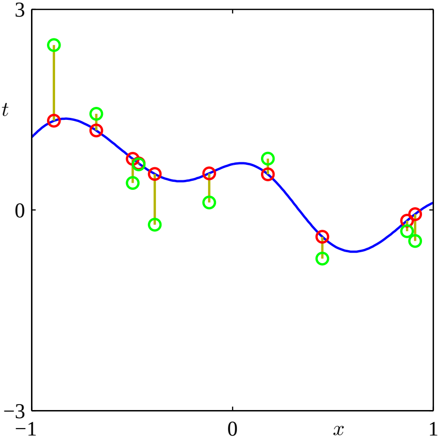
$$p(\mathbf{t}|\mathbf{y}) = \prod_{n=1}^{N} p(t_n|y_n) = \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I})$$

- But $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})$, so linear Gaussian model:

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

with $C_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) + \beta^{-1}\delta_{nm}$

Illustration



Predictive distribution

- Consider prediction $\hat{t}$ at $\hat{\mathbf{x}}$
- Joint distribution of $\mathbf{t}_{N+1} = (t_1, \ldots, t_N, \hat{t})$ is Gaussian, $\mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$
- Covariance matrix in block form:

$$\mathbf{C}_{N+1} = \begin{pmatrix} \mathbf{C} & \mathbf{k} \\ \mathbf{k}^{\mathrm{T}} & c \end{pmatrix}$$

  where $c = k(\hat{\mathbf{x}}, \hat{\mathbf{x}}) + \beta^{-1}$ and $\mathbf{k}$ has elements $k(\mathbf{x}_n, \hat{\mathbf{x}})$
- From results for conditional Gaussians, predictive distribution is also Gaussian,

$$p(\hat{t}|\mathbf{t}) = \mathcal{N}(\hat{t}|\mathbf{k}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t}, c - \mathbf{k}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{k})$$

- That's it! No integrals over $\mathbf{w}$ etc.
- Computational cost dominated by matrix inverse, $O(N^3)$

Illustration for $N = 1$

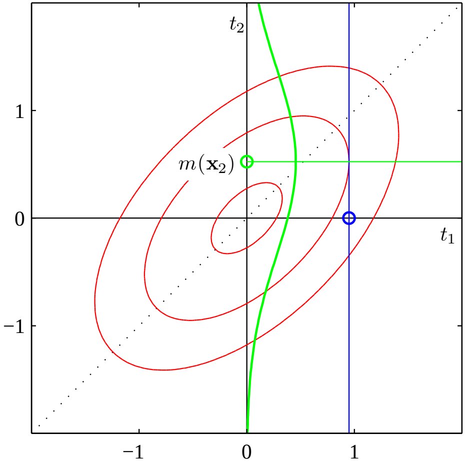
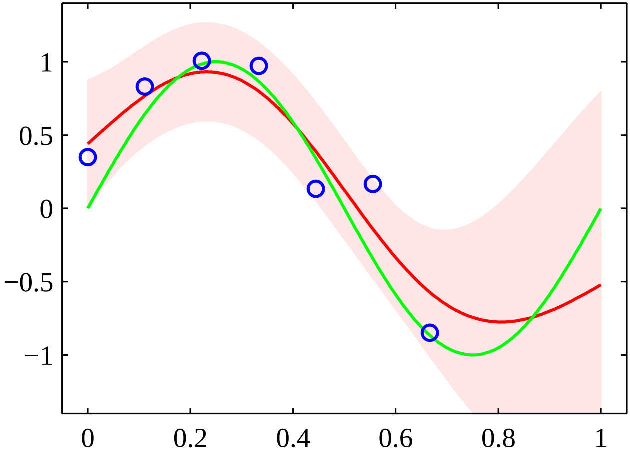


Illustration for sin dataset



Comparison w. (parametric) Bayesian linear regression

- Previously, used prior on weights $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$ and noise model $p(t|\mathbf{x}, \mathbf{w}) = \mathcal{N}(t|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \beta^{-1})$
- Found Gaussian posterior $p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$ with

$$\mathbf{m}_N = \beta\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}, \qquad \mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

- Predictive distribution
  $p(\hat{t}|\hat{\mathbf{x}}, \mathbf{t}) = \mathcal{N}(\hat{t}|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\hat{\mathbf{x}}), \beta^{-1} + \boldsymbol{\phi}(\hat{\mathbf{x}})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\hat{\mathbf{x}}))$
- Should agree with result from GP regression with kernel $k(\mathbf{x}, \mathbf{x}') = \alpha^{-1}\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')$

Marginal likelihood

- Hyperparameters $\boldsymbol{\theta}$: noise level β^{-1} and any kernel parameters like σ^2
- Determine as before by maximizing marginal likelihood $p(\mathbf{t}|\boldsymbol{\theta})$
- Easy – we already know this:

$$\ln p(\mathbf{t}|\boldsymbol{\theta}) = \ln \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C}) = -\frac{1}{2}\ln|\mathbf{C}| - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{t} - \frac{N}{2}\ln(2\pi)$$

- Again, no $\mathbf{w}$-integrals
- Can optimize numerically (generally multiple local maxima)

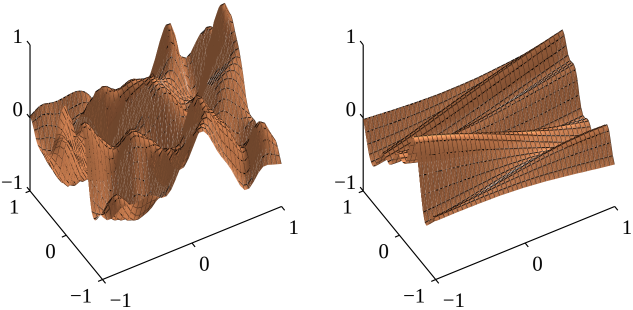
Automatic relevance determination (ARD)

- Can generalize from RBF kernel to

$$k(\mathbf{x}, \mathbf{x}') = \theta_0 \exp\left[-\frac{1}{2}\sum_{i=1}^{D} \eta_i (x_i - x_i')^2\right]$$

- Product of valid kernels, so also valid
- $\eta_i \equiv 1/\sigma_i^2$, so small η_i corresponds to large lengthscale σ_i
- Function $y(\mathbf{x})$ then varies little when x_i is changed
  $\Rightarrow$ input direction i largely irrelevant
- Setting the η_i by maximizing marginal likelihood automatically determines how relevant different input space directions are

Effect of varying η_2



Gaussian processes

1. Discriminative classification
2. Bayesian logistic regression & Laplace approximation
3. Generative classification

Setup for GP classification

- Consider binary class labels $t \in \{0, 1\}$
- Latent function $a(\mathbf{x})$: put GP prior on this
- Likelihood via activation function:

$$p(t|a, \mathbf{x}) = \sigma(a(\mathbf{x}))^t [1 - \sigma(a(\mathbf{x}))]^{1-t}$$

- Predictive distribution: write $\hat{a} = a(\hat{\mathbf{x}})$, $a_n = a(\mathbf{x}_n)$, then

$$p(\hat{t}|\mathbf{t}) = \int p(\hat{t}|\hat{a}) p(\hat{a}|\mathbf{t}) d\hat{a}$$

- Posterior distribution of $\hat{a}$ is, with $\mathbf{a} = (a_1, \ldots, a_N)$

$$p(\hat{a}|\mathbf{t}) = \int p(\hat{a}|\mathbf{a}) p(\mathbf{a}|\mathbf{t}) d\mathbf{a}$$

- Need to approximate $p(\mathbf{a}|\mathbf{t})$, e.g. Laplace approximation