Gaussian Processes in Machine Learning

November 17, 2011

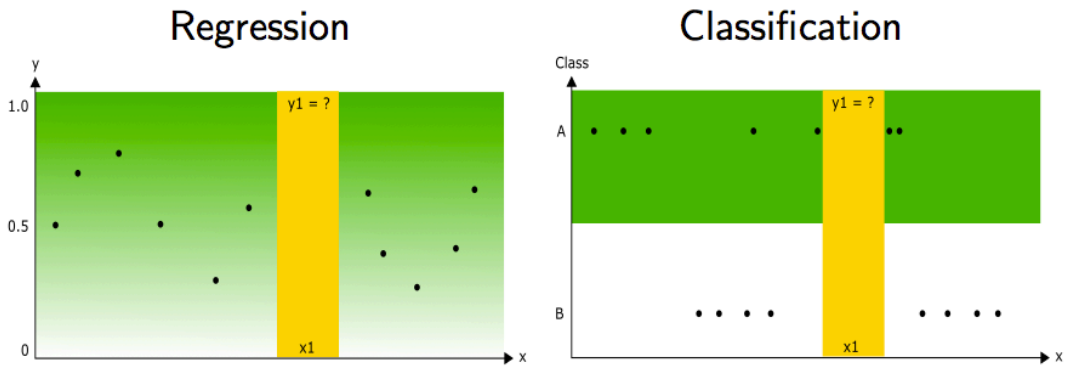
CharmGil Hong

Agenda

- ***Motivation***
- ***GP*** : How does it make sense?
- ***Prior*** : Defining a GP
  - More about Mean and Covariance Functions
- ***Posterior*** : Conditioning with Observations
  - Regression / Classification with GP
- ***& More*** : Real World Applications

Common Problems

- Supervised learning:



  - How to choose a model?
  - How to fit a model to data?

Solutions

- Parametric approaches
  - Polynomials
  - Neural networks
  - Support Vector Machines
- Non-parametric approaches
  - k-NN
  - Gaussian processes

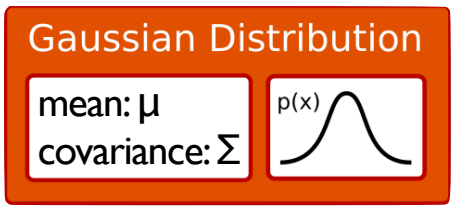
Gaussian Process

Gaussian Process

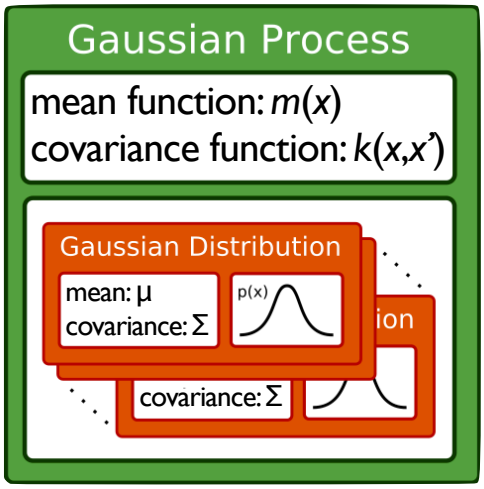
- *Definition* A collection of random variables, any finite number of which have (consistent) Gaussian distribution.
  - A generalization of a multivariate Gaussian distribution to infinitely many variables.
  - A Gaussian process defines a distribution over functions.
    - infinite long vector $\simeq$ function

GD and GP

- vs. Gaussian Distribution



$$X \sim \mathcal{G}(\mu,\Sigma)$$



$$f = (f_1, \ldots f_n)^T \sim \mathcal{GP}(m(x),k(x,x'))$$

GD and GP

- A GD is a distribution over **variables**.
  - It is fully specified by a mean vector and a covariance matrix: $x \sim \mathcal{G}(\mu, \Sigma)$.
  - The position i of x_i is the index in $\boldsymbol{x}$.
- A GP is a distribution over **functions**.
  - It is fully specified by a mean function and a covariance function: $f \sim \mathcal{GP}(m,k)$.
  - The argument x is the index of $f(x)$.

Handling Infinite Dimensionality

- A GP is an infinite dimensional object.
- However, it turns out that we only need to deal with finite dimensional objects.
  - The marginalization property:
    - Recall: $p(x) = \int p(x,y)dy$
    - For Gaussians:

$$p(x,y) = \mathcal{G}\left(\begin{bmatrix} a \\ b \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\right) \Rightarrow p(x) = \mathcal{G}(a,A)$$

Prior

Defining a GP

- *Definition* $P(f)$ is a Gaussian process if for any finite subset $\{x_1,...,x_n\} \subset X$, the marginal distribution over that finite subset $P(f)$ has a multivariate Gaussian distribution.
- Let $f = (f(x_1),...,f(x_n))$ be an N-dimensional vector of function values evaluated at N points $x_i \in X$.
- Again, $f(x_i)$ is now a random variable and each x_i is the index (cf. in GD, x_i is a random variable and i is the index).

Defining a GP

- How to define a GP?
  - Choose a form for the mean function.
  - Choose a form for the covariance function.
- Recall: a GP is fully specified by a mean function and a covariance function: $f \sim GP(\, m, k \,)$.

Mean and Covariance Fn

- Any functions can be a mean function and a covariance* function. (*will be revisited)
- Usually,
  - The mean function is usually defined to be zero.
  - Several covariance functions have been used in the literature, but the predominant choice is a squared exponential (SE).

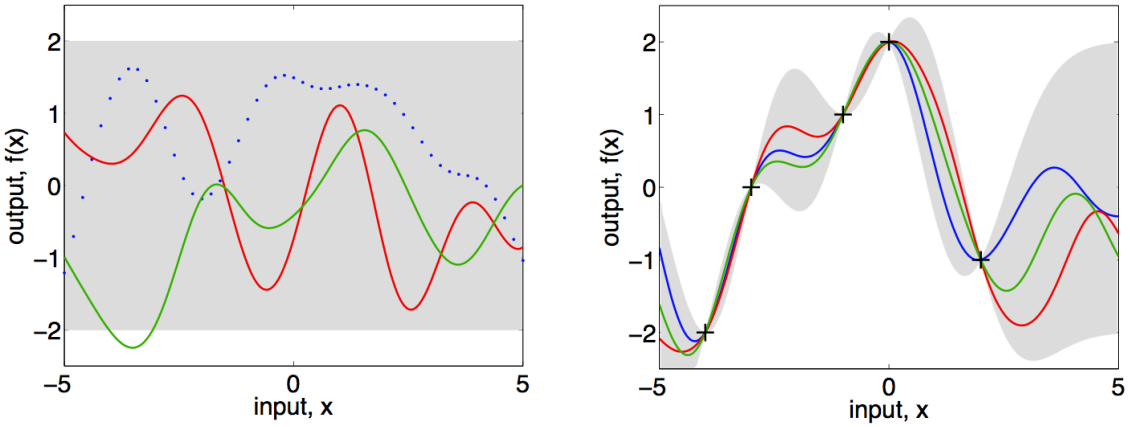
Mean and Covariance Fn

- Squared Exponential (SE)
  - $k(x_i,x_j) = v_0 exp\left\{-\frac{1}{2}\sum_{m=1}^{d} l_m(x_i^m - x_j^m)^2\right\} + v_1 + v_2\delta_{ij}$

    where x_i^m is the m-th element of x_i.
  - SE depends on hyperparameters v_o, v_1, v_2, and l_m.
    - l_m: characteristic length-scale.
    - v_o: overall vertical scale of variation of the latent value.
    - v_1: overall bias of the latent values.
    - v_2: latent noise variance.

Meaning of the Covariance Fn

- The covariance function defines how smoothly the (latent) function f varies from a given x.
- The data points "anchor" the function f at specific x locations.



Properties of the Covariance Fn

- Only restriction is that it must be positive semi-definite (PSD).
  - *Theorem* If k, k_1, and k_2 are PSD, then the following are also PSD:
    1. $\alpha k(x,y) \quad \alpha \geq 0$
    2. $k_1(x,y) + k_2(x,y)$
    3. $k_1(x,y)k_2(x,y)$
    4. $P(k(x,y))$, *where* $P(x)$ *with non-negative coefficients*

Properties of the Covariance Fn

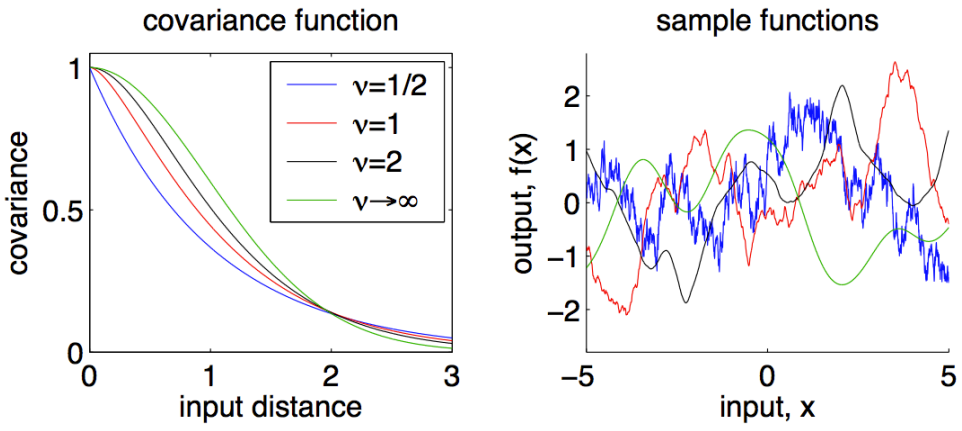
- Only restriction is that it must be positive semi-definite (PSD).
  - *Theorem* If k, k_1, and k_2 are PSD, then the following are also PSD:

5. $exp(\ k(x,y)\ )$
6. $f(x)k(x,y)f(y)$
7. $k(\psi(x), \psi(y))$

Covariance Fn Examples

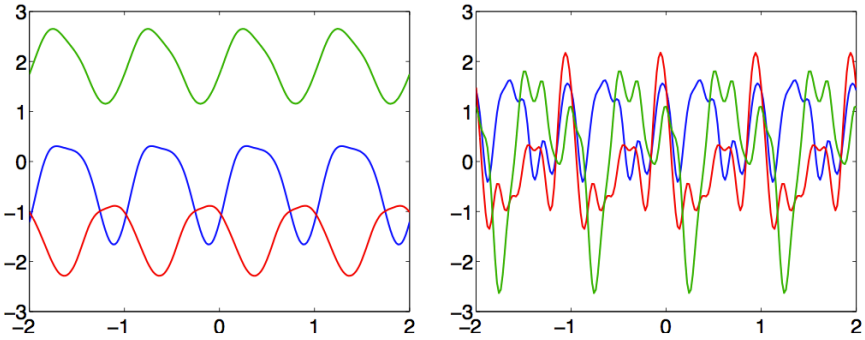
- Matérn Covariance Function
  - $$k(\mathbf{x}, \mathbf{x}') = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left[\frac{\sqrt{2\nu}}{\ell}|\mathbf{x}-\mathbf{x}'|\right]^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}}{\ell}|\mathbf{x}-\mathbf{x}'|\right),$$

    where Kv is the modified *Bessel* function of second kind of order v.



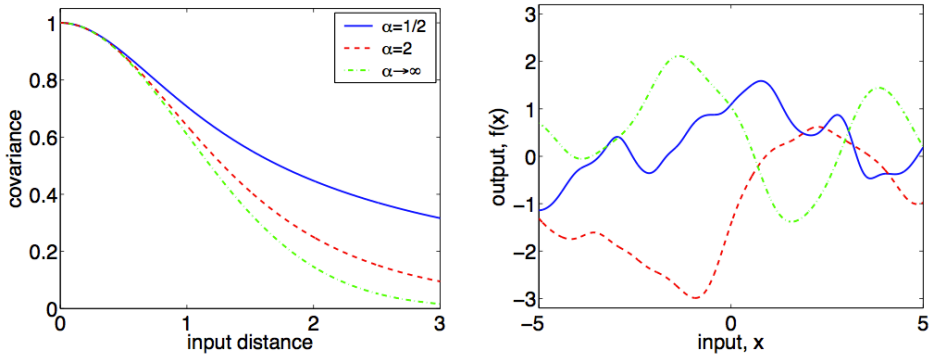
Covariance Fn Examples

- Periodic, smooth functions
  - $k_{periodic}(x,x') = exp(-2\ sin^2(\pi(x-x'))/l^2)$



Covariance Fn Examples

- Rational Quadratic
  - $k_{\mathrm{RQ}}(r) = \left(1 + \frac{r^2}{2\alpha\ell^2}\right)^{-\alpha}$



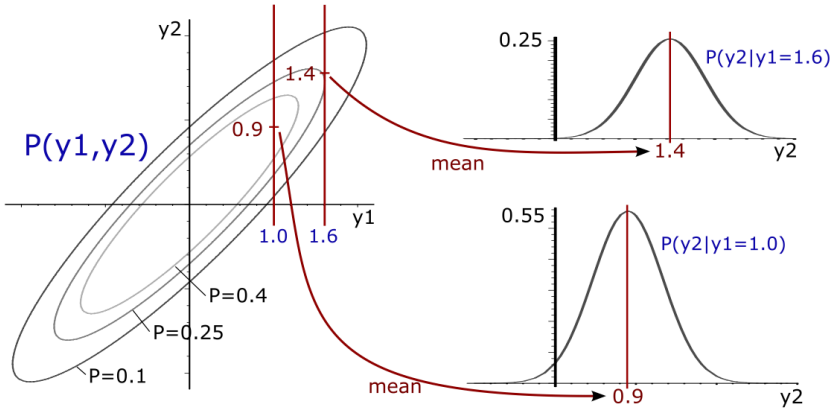
Bayesian Regression

- Then, what can we do with Gaussian processes?
  - A GP can be a prior of a Bayesian regression problems.
  - A GP prior actually offers rather simpler solution!

Posterior

Conditional Distribution $P(y_2|y_1)$

- Let say we have the covariance matrix K and the value of y_1. Then the posterior distribution $P(y_2|y_1)$ is also a Gaussian.
- Our goal is to determine the mean and the corresponding variance of y_2 given y_1.



Conditional Distribution $P(y_2|y_1)$

- $P(y_2|y_1,K)$

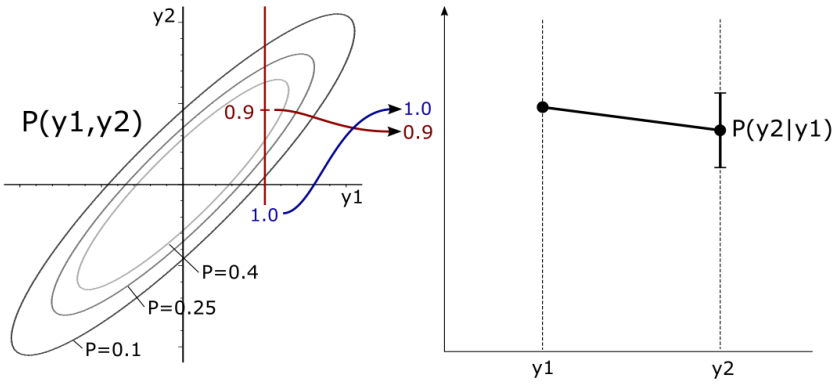
$$
\begin{aligned}
&= \frac{P(y_1, y_2|K)}{P(y_1|K)} && (1)\\
&\propto exp - \frac{1}{2}\left\{ \begin{pmatrix} y_1 & y_2 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\} && (2)\\
&= exp - \frac{1}{2}\left\{ y_1^2 a + 2y_1 y_2 b + y_2^2 c \right\} && (3)\\
&\propto exp - \frac{1}{2}\left\{ 2y_1 y_2 b + y_2^2 c \right\} && (4)\\
&= exp - \frac{1}{2}\left\{ (y_2^2 + 2y_2 y_1 \frac{b}{c})c \right\} && (5)\\
&\propto exp - \frac{1}{2}\left\{ (y_2^2 + 2y_2 y_1 \frac{b}{c} + y_1^2 \frac{b^2}{c^2})c \right\} && (6)\\
&= exp - \frac{1}{2}\left\{ ((y_2 + y_1 \frac{b}{c})^2)c \right\} && (7)\\
&= exp - \frac{1}{2}\left\{ \frac{(y_2 - (-y_1 \frac{b}{c}))^2}{1/c} \right\} && (8)
\end{aligned}
$$

Conditional Distribution $P(y_2|y_1)$

- Let K = $\begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ and $y_1 = 1.0$,

  then we get $K^{-1} = \begin{pmatrix} 5.26 & -4.74 \\ -4.74 & 5.26 \end{pmatrix} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$.

  Now we are able to obtain $y_2 = N(0.9, 0.19)$



Expending to Vectors

- $P(y_2|y_1,K) = \frac{P(y_1, y_2|K)}{P(y_1|K)}$

$$\propto exp - \frac{1}{2}\left\{ (y_1^T \quad y_2^T) \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\}$$

$$= exp - \frac{1}{2}\left\{ y_1^T A y_1 + y_2^T B^T y_1 + y_1 T B y_2 + y_2^T C y_2 \right\}$$

$$\propto exp - \frac{1}{2}\left\{ y_2^T C y_2 + y_2^T B^T y_1 + y_1^T B y_2 \right\}$$

$$\propto exp - \frac{1}{2}\left\{ y_2^T C y_2 + y_2^T B^T y_1 + y_1^T B y_2 + y_1^T B C^{-1} B^T y_1 \right\}$$

$$= exp - \frac{1}{2}\left\{ (y_2^T C + y_1^T B)(y_2 + C^{-1} B^T y_1) \right\}$$

$$= exp - \frac{1}{2}\left\{ (y_2^T + y_1^T B C^{-1}) C (y_2 + C^{-1} B^T y_1) \right\}$$

$$= exp - \frac{1}{2}\left\{ (y_2 - (-C^{-1} B^T y_1)) C (y_2 - (-C^{-1} B^T y_1)) \right\}$$

GP for Regression

- Goal: Predict the real-values output y^* for anew input value x^*.
- Given: Training data $D = \{(\boldsymbol{x}_i, y_i),\ i=1,\ldots,N\}$.
- Model: $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$.
  - Prior: $f \sim \mathcal{GP}(\cdot|m,k)$
  - Noise: $\varepsilon_i \sim \mathcal{G}(\cdot|0,\sigma^2)$

GP for Regression

- Model: $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$.
  - Prior: $f \sim \mathcal{GP}(\cdot|m,k)$, Noise: $\varepsilon_i \sim \mathcal{G}(\cdot|0,\sigma^2)$
- The covariance function k depends on a set of hyperparameters $\boldsymbol{w}$.
  - Recall: $k(x_i,x_j) = v_0 exp\left\{-\frac{1}{2}\sum_{m=1}^{d} l_m(x_i^m - x_j^m)^2\right\} + v_1 + v_2\delta_{ij}$
- The problem of learning with GP is exactly the problem of learning the hyperparameters.
  - Once the hyperparameters are learned, inference can be performed.

GP for Regression

- Maximum likelihood (method 1)
  - Gaussian likelihood:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M) \propto \prod_i \exp(-\tfrac{1}{2}(y_i - f(x_i))^2/\sigma^2_{\text{noise}})$$

  - Maximize the likelihood:

$$\mathbf{w}_{\text{ML}} = \underset{\mathbf{w}}{\text{argmax}}\, p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M)$$

  - Make predictions, by plugging in the ML estimate:

$$p(y^*|x^*, \mathbf{w}_{\text{ML}}, M)$$

GP for Regression

- Bayesian Inference (method 2)
  - Gaussian likelihood:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M) \propto \prod_i \exp(-\tfrac{1}{2}(y_i - f(x_i))^2/\sigma^2_{\text{noise}})$$

  - Parameter prior:

$$p(\mathbf{w}|M)$$

  - Posterior by Bayes rule:

$$p(\mathbf{w}|\mathbf{x}, \mathbf{y}, M) = \frac{p(\mathbf{w}|M)p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M)}{p(\mathbf{y}|\mathbf{x}, M)}$$

  - Make predictions:

$$p(y^*|x^*, \mathbf{x}, \mathbf{y}, M) = \int p(y^*|\mathbf{w}, x^*, M)p(\mathbf{w}|\mathbf{x}, \mathbf{y}, M)d\mathbf{w}$$

GP for Regression

- Non-parametric GP models (method 3)
  - In this method, the "parameters" is the function itself!
  - Gaussian likelihood: $p(\mathbf{y}|\mathbf{x}, \mathbf{w}, M) \sim \mathcal{N}(\mathbf{f}, \sigma^2_{\text{noise}} I)$
  - Gaussian process prior:

$$p(f(x)|M) \sim \mathcal{GP}\big(m(x) \equiv 0,\ k(x, x')\big)$$

  - Gaussian process posterior:

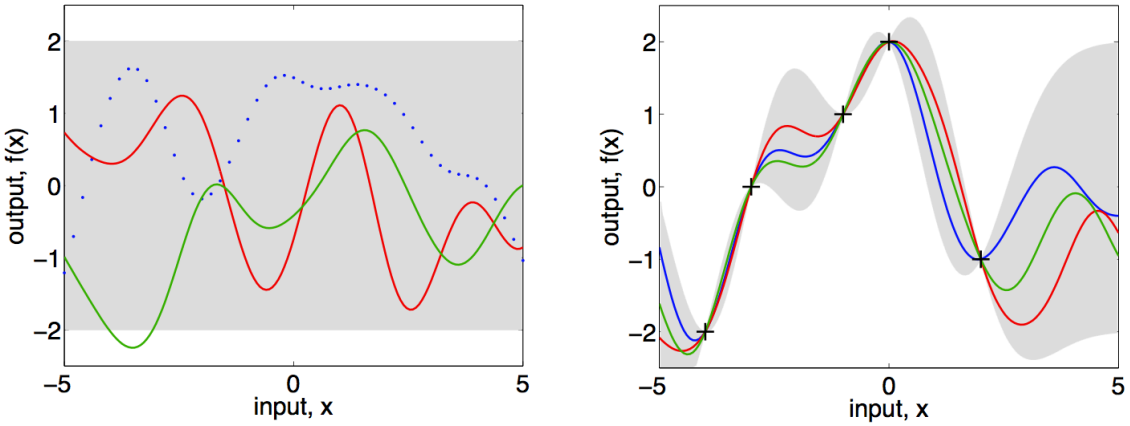
$$p(f(x)|\mathbf{x}, \mathbf{y}, M) \sim \mathcal{GP}\big(m_{\text{post}}(x) = k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y},\ k_{\text{post}}(x, x') = k(x, x') - k(x, \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma^2_{\text{noise}} I]^{-1} k(\mathbf{x}, x')\big)$$

GP for Regression

- Non-parametric GP models (method 3)
  - Gaussian predictive distribution:

$$p(y^*|x^*, \mathbf{x}, \mathbf{y}, M) \sim \mathcal{N}\big(\mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1}\mathbf{y},\ k(x^*, x^*) + \sigma^2_{\text{noise}} - \mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1}\mathbf{k}(x^*, \mathbf{x})\big)$$

GP for Regression



- Predictive Distribution:

$$p(y^*|x^*, \mathbf{x}, \mathbf{y}, M)$$
$$\sim \mathcal{N}\big(\mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1} \mathbf{y},$$
$$k(x^*, x^*) + \sigma^2_{\text{noise}} - \mathbf{k}(x^*, \mathbf{x})^\top [K + \sigma^2_{\text{noise}} I]^{-1} \mathbf{k}(x^*, \mathbf{x})\big)$$

GP for Classification

- Goal: Predict the label y^* of a new input value x^*.
  - $y \in \{-1, 1\}$.
- $L_{\text{prediction}} = p(y_i|x_i) = \sigma(f(x_i))$, where σ is a sigmoid transformation (e.g., logistic function or cumulative distribution function of standard normal distribution).
- Marginal likelihood:
  - $P(f|D,w) = \int \sigma(f(x_i)) P(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{w}) df$.
  - This integral is a product of sigmoids (likelihood) multiplied by a Gaussian (prior), and is therefore intractable.

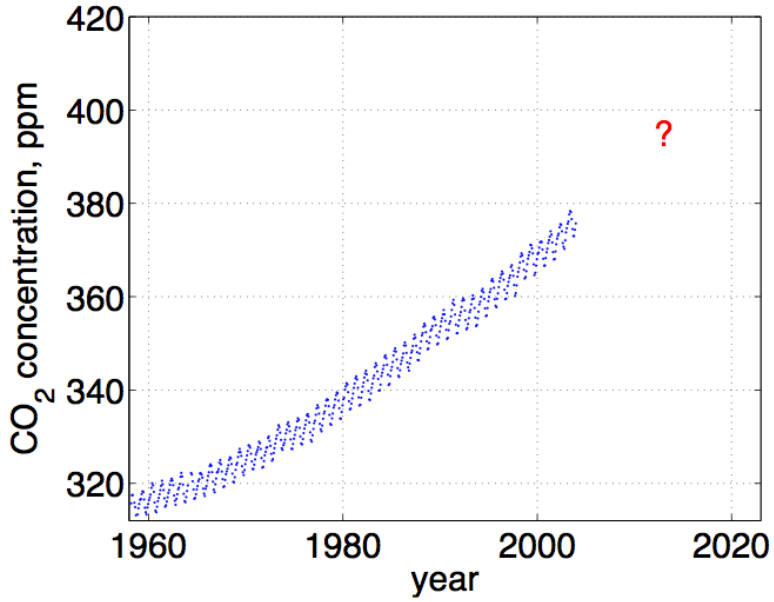
Tractability of the Posterior

- In regression, a Gaussian likelihood and the Gaussian process prior result in a tractable posterior.
- In classification, however, the posterior $P(f|D,\theta)$ is intractable, since it involves an integral that is the product of a Gaussian and a product of sigmoids.
  - Approximation is required.
  - e.g. Laplace approximation, Expectation-Propagation, Variational method, MCMC sampling.

& More

Applications

- CO_2 prediction problem



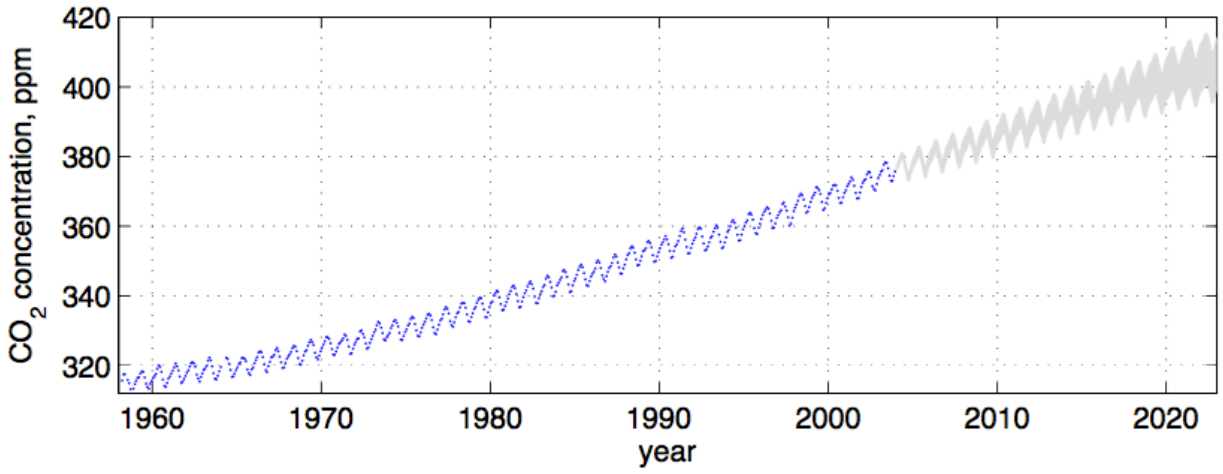
Applications

- Build a covariance function:
  - long-term smooth trend (squared exponential)
    $k_1(x, x') = \theta_1^2 \exp(-(x - x')^2/\theta_2^2)$,
  - seasonal trend (quasi-periodic smooth)
    $k_2(x, x') = \theta_3^2 \exp\left(-2\sin^2(\pi(x - x'))/\theta_5^2\right) \times \exp\left(-\frac{1}{2}(x - x')^2/\theta_4^2\right)$,
  - short- and medium-term anomaly (rational quadratic)
    $k_3(x, x') = \theta_6^2\left(1 + \frac{(x-x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8}$
  - noise (independent Gaussian, and dependent)
    $k_4(x, x') = \theta_9^2 \exp\left(-\frac{(x-x')^2}{2\theta_{10}^2}\right) + \theta_{11}^2\delta_{xx'}$.

$$k(x, x') = k_1(x, x') + k_2(x, x') + k_3(x, x') + k_4(x, x')$$

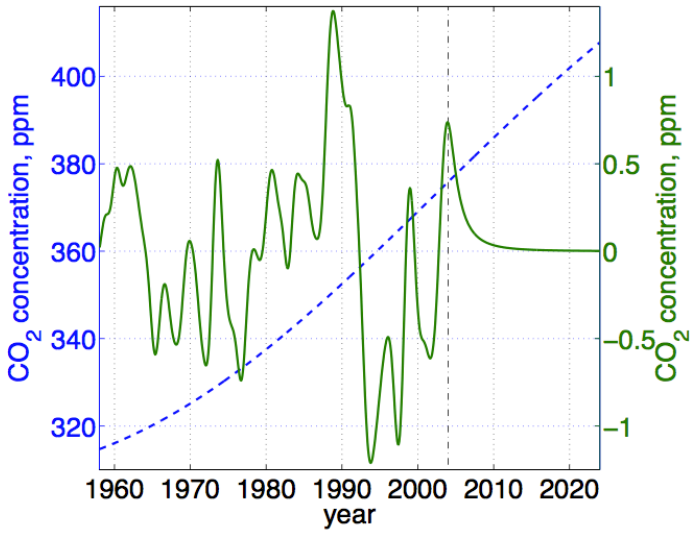
Applications

- CO_2 predictions



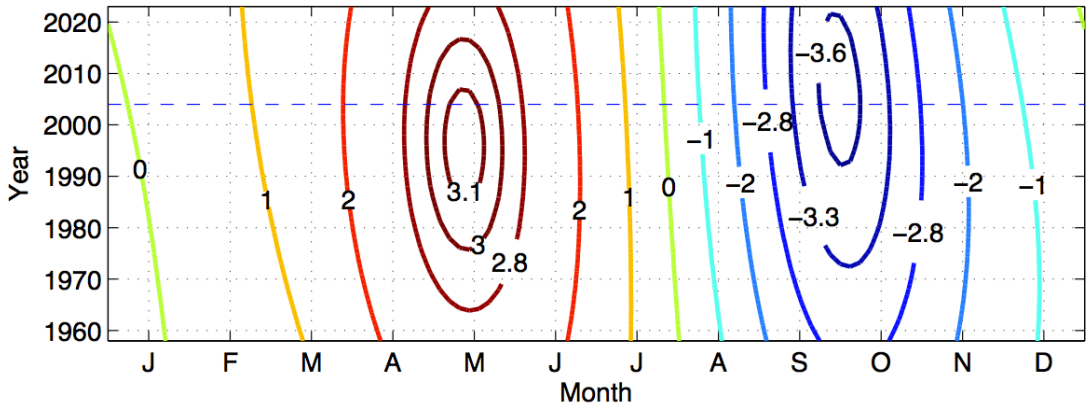
Applications

- Long-/medium-/mean predictions



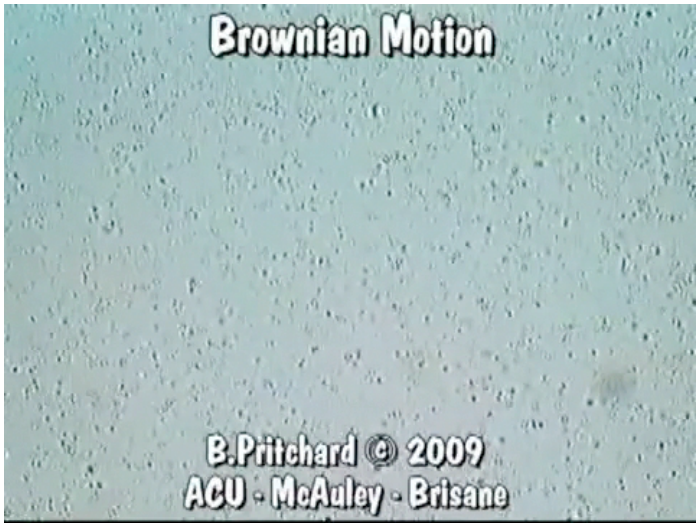
Applications

- Mean Seasonal predictions



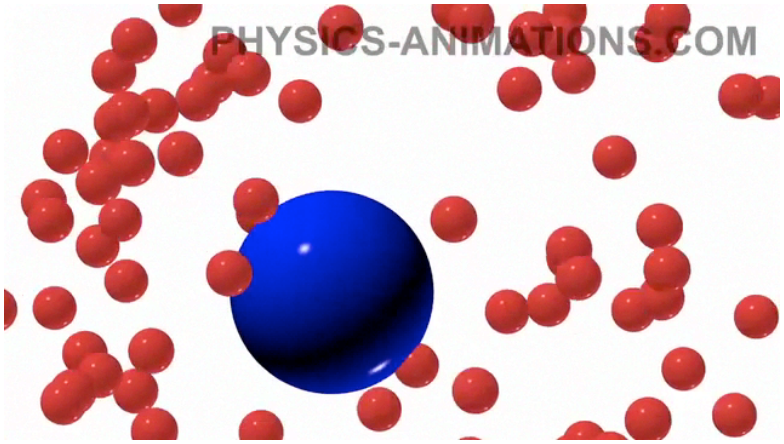
Applications

- Molecule movement modeling (using Matérn Covariance Function)



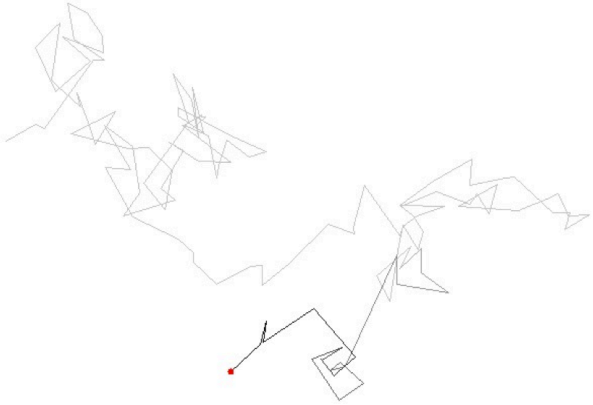
Applications

- Molecule movement modeling (using Matérn Covariance Function)



Applications

- Molecule movement modeling (using Matérn Covariance Function)
  - $X(t) - X(t') \sim \mathcal{G}(o, t-t')$
  - $X(t) \sim \mathcal{GP}(o, min(t,t'))$



Applications

- More Applications
  - Handwriting recognition
  - Determining trustworthiness of bank clients
  - Generating music playlists
  - Articulated body tracking

Summary

- Gaussian processes are non-parametric.
- A Gaussian process is fully specified by a mean function and covariance function.
- The problem of learning with Gaussian processes is exactly the problem of learning the hyperparameters of the covariance function.
- Basic rules of multivariate Gaussian distribution govern manipulation of the Gaussian process after a finite number of data points is observed.

Summary

- GPs offer a more general approach than standard logistic regression.
- GPs can be used in a Bayesian setting where the GP is a prior on the function.
- GPs can handle the case in which data is available in (multiple) different forms, as long as we can define an appropriate covariance function for each data type.

Drawbacks

- The basic complexity of Gaussian process is $O(N^3)$ where N is the number of data points, due to the inversion of an N x N matrix.
  - Practical limit is said to be $N \approx 1000$ or fewer.
- Classification results intractable posteriors.
  - Approximation must be employed.

References

- Papers
  - C. Rasmussen, “Gaussian Processes in Machine Learning,” 2003.
  - E. Ebden, “Gaussian Processes for Regression: A Quick Introduction,” 2008.
  - E. Ebden, “Gaussian Processes for Classification: A Quick Introduction,” 2008.
- Slides
  - D. Williams, “Gaussian Process,” 2006.
  - A. Geiger, “An introduction to Gaussian Processes, (scaled) GPLVMs, (balanced) GPDMs and their applications to 3D people tracking,” 2007.
  - C. Rasmussen, “Learning with Gaussian Processes,” 2008.