Gaussian Processes

A general overview plus discussion of the paper

“Assessing Approximations for Gaussian Process Classification”

by Malte Kuss and Carl Edward Rasmussen (from NIPS 2005)

DUKE UNIVERSITY

Machine Learning Research Group - Paper Discussion

January 27, 2006

Presented by David P. Williams

Outline

- Gaussian Processes
- Paper Discussion

Relationship to Logistic Regression

- In logistic regression, the input to the sigmoid function is $f = \boldsymbol{w}^T \boldsymbol{x}$ or $f = \boldsymbol{w}^T \phi(\boldsymbol{x})$, where $\boldsymbol{w}$ are (classifier) parameters.
- A Gaussian process places a prior on the space of functions f directly, *without parameterizing* f.
- Therefore, Gaussian processes are non-parametric (*e.g.*, no $\boldsymbol{w}$ used explicitly).

Why Gaussian Processes?

- GPs are more general than standard logistic regression because the form of the classifier is not limited by a parametric form.
- GPs can be used in a Bayesian setting where the GP is a prior on the functions.
- 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.
  - standard vector data
  - sequences (*e.g.*, as in biological data)
  - images
  - ...

Gaussian Distributions and Gaussian Processes

- A ***Gaussian distribution*** is a distribution over ***vectors***.
- It is fully specified by a mean and a covariance: $\boldsymbol{x} \sim \mathcal{G}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- The **position** of the random variables x_i in the vector plays the role of the index.

- A ***Gaussian process*** 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** $\boldsymbol{x}$ of the random function $f(\boldsymbol{x})$ plays the role of the index.

Handling Infinite Dimensional Objects

- A Gaussian Process (GP) is an infinite dimensional object.
- However, it turns out that we will only ever need to work with finite dimensional objects. (Why?)
- *Definition*: A Gaussian process is a collection of random variables, any finite number of which have joint Gaussian distributions.
- Conditioning the GP on the function values at observed values of x will be key.
- In our work, x would correspond to data points (*e.g.*, features).

Definition

- Let $\boldsymbol{f} = (f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N))$ be an N-dimensional vector of function values evaluated at N points $\boldsymbol{x}_i \in \mathcal{X}$.
- Note that $\boldsymbol{f}$ is a random variable.
- *Definition*: $P(f)$ is a Gaussian process if for *any* finite subset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathcal{X}$, the marginal distribution over that finite subset $P(\boldsymbol{f})$ has a multivariate Gaussian distribution.

Defining a Gaussian Process

- So how does one define a Gaussian process?
- Recall that a GP is fully specified by a mean function and a covariance function: $f \sim \mathcal{GP}(m, k)$.
- **The mean function and covariance function drive the entire GP.**
- We need two things to define our GP:
  - We need to *choose* a form for the mean function.
  - We need to *choose* a form for the covariance function.

Mean Function and Covariance Function

- 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” (a.k.a. Gaussian or RBF) covariance function of the form

$$K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = v_0 \exp\left\{ -\frac{1}{2} \sum_{m=1}^{d} \ell_m (x_i^m - x_j^m)^2 \right\} + v_1 + v_2 \delta(i, j)$$

  where x_i^m is the m-th element of $\boldsymbol{x}_i$.
- Note that this covariance function depends on hyperparameters v_0, v_1, v_2, and ℓ_m.

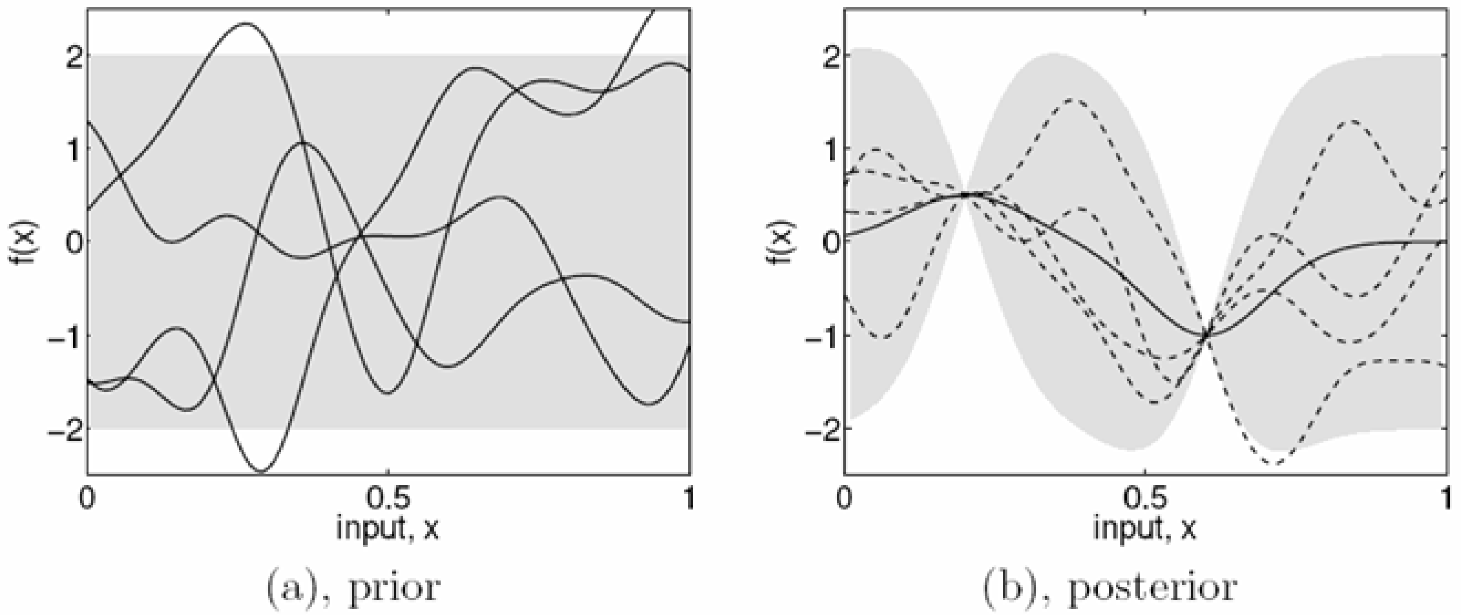
Hyperparameters

- $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = v_0 \exp\left\{-\frac{1}{2}\sum_{m=1}^{d} \frac{(x_i^m - x_j^m)^2}{\ell_m}\right\} + v_1 + v_2\delta(i,j)$
- ℓ_m: characteristic length-scale
  - roughly the distance you must move in input space before the function value can change significantly.
  - short length-scales mean the error bars (*i.e.*, predictive variance) can grow rapidly away from the data points.
  - large length-scales imply irrelevant features (function value would be constant function of that feature input).
- v_0: overall vertical scale of variation of the latent value.
- v_1: overall bias of the latent values from zero mean; akin to a basis vector of ones (bias offset term) in logistic regression.
- v_2: latent noise variance; “jitter” that makes matrix computations better conditioned.

Meaning of the Covariance Function

- What does the covariance function **represent**? What does it imply?
- The covariance function defines how smoothly the (latent) function f varies from a given $\boldsymbol{x}$.
- The data points “anchor” the function f at specific $\boldsymbol{x}$ locations. (see next slide)

Implication of the Covariance Function



(a), prior

(b), posterior

Figure 1.1: Panel (a) shows four samples drawn from the prior distribution. Panel (b) shows the situation after two datapoints have been observed. The mean prediction is shown as the solid line and four samples from the posterior are shown as dashed lines. In both plots the shaded region denotes twice the standard deviation at each input value x.

Properties of the Covariance Function

- The only technical restriction on the covariance function is that it must be positive semi-definite.
- It can be non-stationary (*e.g.*, the length-scale may depend on the values of $\boldsymbol{x}$).
- Covariance function can be the sum (or product or linear combination) of other covariance functions

  e.g., can use a different covariance function for each unique sensor modality or data type (vector, sequence, image data).

Mean and Covariance

- Once we have a group of $\boldsymbol{x}$ values (*i.e.*, data points), we can compute the mean vector and covariance matrix for the GP.
- Recall that the argument $\boldsymbol{x}$ of the random function $f(\boldsymbol{x})$ plays the role of the index.
- Note that for N observed data points, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, the mean vector $\boldsymbol{m}$ will be an N-element column vector, and the covariance matrix $\mathbf{K}$ will be an $N \times N$ matrix. (Note the computational issues this may potentially raise.)
- Thus, once we (1) choose the form of the mean function and covariance function, and (2) observe some data, we will have a multivariate Gaussian: $\boldsymbol{f} \sim \mathcal{G}(\boldsymbol{m}, \mathbf{K})$.
- Equivalently: $p(\boldsymbol{f}) = p(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N)) = \frac{1}{(2\pi)^{N/2}|\mathbf{K}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{f} - \boldsymbol{m})^T \mathbf{K}^{-1} (\boldsymbol{f} - \boldsymbol{m})\right\}$

Conditioning a Gaussian Distribution

- Once the mean function and covariance function are defined, GPs are ruled by basic probability applied to multivariate Gaussian *distributions*.
- For the joint distribution defined as

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{G}\left(\begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{m}_* \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{bmatrix} \right),$$

the conditional distribution will be

$$\boldsymbol{f}_*|\boldsymbol{f} \sim \mathcal{G}\left(\boldsymbol{m}_* + \mathbf{K}_*^T\mathbf{K}^{-1}(\boldsymbol{f} - \boldsymbol{m}), \mathbf{K}_{**} - \mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}_*\right)$$

Aside: Relationship to Dirichlet Processes

- A ***Gaussian process*** defines a distribution over **functions**: $f \sim \mathcal{GP}(m, k)$
- GPs can be viewed as infinite-dimensional Gaussian distributions.

- A ***Dirichlet process*** defines a distribution over **distributions**: $G \sim \mathcal{DP}(G_0, \alpha_0)$
- DPs can be viewed as infinite-dimensional Dirichlet distributions.

- Note that both f and G are infinite-dimensional objects.

Midway Summary

- Gaussian processes are non-parametric.
- A Gaussian process is a collection of random variables, any finite number of which have joint Gaussian distributions.
- A Gaussian process is fully specified by a mean function and a covariance function.
- Basic rules of multivariate Gaussian distributions govern manipulation of the Gaussian process after a finite number of data points is observed.

Uses of Gaussian Processes

- So how/when does one use a GP?
- GPs are used in regression and classification.

Gaussian Processes for Regression

- Goal: Predict the real-valued output y' for a new input value $\boldsymbol{x}'$.
- Given: training data $D = \{(\boldsymbol{x}_i, y_i), i = 1, \ldots, N\}$
- Model: $y_i = f(\boldsymbol{x}_i) + \epsilon_i$
  - Noise: $\epsilon_i \sim \mathcal{G}(\cdot|0, \sigma^2)$
  - Prior: $f \sim \mathcal{GP}(\cdot|0, k)$
- Covariance function k depends on a set of hyperparameters $\boldsymbol{\theta}$.
- Prior on f is a GP, and likelihood is Gaussian, so posterior on f is also a GP: $P(f|D, \boldsymbol{\theta}) \propto P(D|f) \times P(f|\boldsymbol{\theta})$.
- Make predictions with: $P(y'|\boldsymbol{x}', D) = \int df\, P(y'|\boldsymbol{x}', f, D) P(f|D, \boldsymbol{\theta})$

Gaussian Processes for Classification

- In classification, $y_i \in \{-1, 1\}$.
- $p(y_i|\boldsymbol{x}_i) = \sigma(f(\boldsymbol{x}_i))$, where σ is a sigmoid transformation (*e.g.*, logistic function or cumulative distribution function of standard normal distribution).
- Marginal likelihood (*i.e.*, evidence) is the integral $\int P(\boldsymbol{y}|\boldsymbol{f})P(\boldsymbol{f}|\mathbf{X}, \boldsymbol{\theta})d\boldsymbol{f}$.
- Integral is a product of sigmoids (likelihood) multiplied by a Gaussian (prior), and is therefore intractable.
    - Recall that in the regression case, the likelihood was a Gaussian, which made the integration tractable.
- Thus, the posterior cannot be computed analytically.
- Some approximation must be employed to obtain an approximate posterior.

Tractability of the Posterior

- In regression, a Gaussian likelihood and the Gaussian process prior result in a tractable posterior.
- In classification, the posterior $P(\boldsymbol{f}|D, \Theta)$ is intractable because it involves an integral that is the product of a Gaussian and a product of sigmoids.
- Several different techniques have been proposed to overcome this obstacle:
    - Laplace approximation [Barber & Williams]
    - Variational methods [Gibbs & MacKay]
    - Expectation-Propagation [Minka & Ghahramani]
    - MCMC sampling [Neal]

Approximation Methods (1)

- Laplace approximation
  - Make a Taylor approximation of the un-normalized log-posterior.
  - Mean $\boldsymbol{m}$ is placed at the mode (MAP).
  - Covariance $\mathbf{A}$ equals the negative inverse Hessian of the log-posterior density at $\boldsymbol{m}$.
- Expectation-Propagation
  - Gaussian approximation to the posterior.
  - Parameters $\boldsymbol{m}$ and $\mathbf{A}$ are found in an iterative scheme by matching the approximate marginal moments of $p(f_i|D,\boldsymbol{\theta})$ by the marginals of the approximation $\mathcal{G}(f_i|\boldsymbol{m}_i,\mathbf{A}_{ii})$.

Approximation Methods (2)

- Variational methods
  - Place lower and upper bounds on the sigmoid function.
  - Optimize the bounds with respect to variational parameters.
- MCMC sampling
  - Obtain samples from the posterior via Gibbs sampling.
  - Becomes exact in the limit of long runs (“gold standard”).

Learning with Gaussian Processes

- Recall that the covariance function depends on several hyperparameters:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = v_0 \exp\left\{ -\frac{1}{2} \sum_{m=1}^{d} \frac{(x_i^m - x_j^m)^2}{\ell_m} \right\} + v_1 + v_2 \delta(i, j)$$

- **The problem of learning with Gaussian processes is exactly the problem of learning these hyperparameters.**
  - Can place (Gamma) priors on the hyperparameters and get posterior distributions of the hyperparameters.
  - Can optimize the hyperparameters directly.
- Once the hyperparameters are decided upon, inference can be performed.

Drawbacks

- In Gaussian process classification, the posterior is intractable, so approximations must be employed.
- The basic complexity of Gaussian processes is $\mathcal{O}(N^3)$ where N is the number of data points, due to the inversion of an $N \times N$ matrix.
    - Limits method to case in which $N \approx 1000$ or fewer.

One Slide Summary of Paper

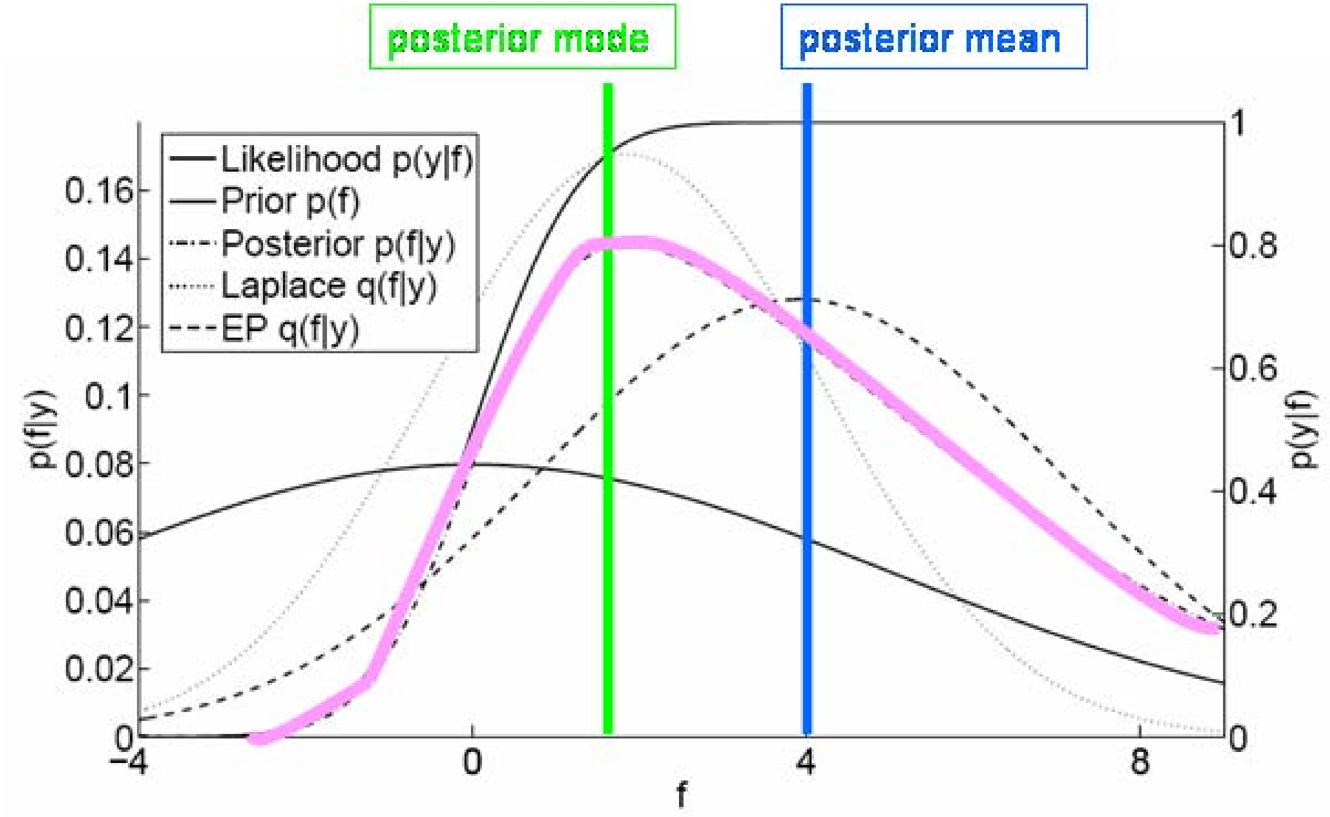
“Assessing Approximations for Gaussian Process Classification” by Malte Kuss and Carl Edward Rasmussen (from NIPS 2005)

- Compares two approximations — Laplace approximation (LA) and Expectation Propagation (EP) — for Gaussian process classification.
- Found EP to be much more accurate than the Laplace approximation.

Approximation Methods

- Both approximations, EP and LA, are based on a Gaussian approximation to the posterior.
- Each approximation uses a different method to find a mean and covariance of the approximate posterior.
  - The Laplace approximation will match the **posterior mode** (by construction).
  - The EP approximation will match the **first two posterior moments**.
- What does this imply?

Implications of Approximations



Structural Properties of the Posterior

- The prior is a correlated N-dimensional Gaussian $\mathcal{G}(\boldsymbol{f}|\mathbf{0},\mathbf{K})$ centered at the origin.
- Each likelihood term $p(y_i|f_i)$ *softly* truncates the half-space from the prior that is incompatible with the observed label.
- Resulting posterior is unimodal and skewed.

Properties of High-Dimensional Gaussians

- In high-dimensional Gaussians, most probability mass is contained in a thin ellipsoidal shell away from the mean.
- Is because the volume grows rapidly with the radius, in high dimensions.
- The mode becomes less representative as the dimension increases.
- For GP classification posterior, mode of the posterior distribution stays close to the origin, but mean moves to the mass of the posterior.
- Therefore, posterior mode and mean are significantly different.

Implications of Laplace Approximation

- Laplace approximation places the mean m in the correct orthant, but too close to the origin.
- The approximated posterior will overlap with regions of very little posterior mass.
- The amplitude of the approximate posterior will be systematically underestimated, leading to overly cautious predictive distributions (*i.e.*, predictions closer to 0.5).
- Authors found experimentally that predictive class probabilities were inaccurate even at *training* locations.

Success of EP Approximation

- Recall that EP matches the approximate marginal moments of the posterior $p(f_i|D, \boldsymbol{\theta})$ by the marginal moments of the approximation $\mathcal{G}(f_i|\boldsymbol{m}_i, \mathbf{A}_{ii})$.
- EP seems to succeed because the marginal distributions of the posterior are well-approximated by Gaussians.
  - Authors justified this claim by experimentally finding that the marginal distribution of a truncated high-dimensional Gaussian was well-approximated by a Gaussian.
  - Laplace approximation still fails because this approximate Gaussian is not necessarily centered near the origin.

Take-Home Message from Paper

- Use EP for approximate inference in Gaussian process classification models when the computational cost of MCMC is prohibitive.
- Laplace approximation is inaccurate and should not be used.

Gaussian Process Take-Home Message

- Gaussian processes are non-parametric.
- A Gaussian process is a collection of random variables, any finite number of which have joint Gaussian distributions.
- A Gaussian process is fully specified by a mean function and a 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 distributions govern manipulation of the Gaussian process after a finite number of data points is observed.

References

- M. Gibbs and D. MacKay, “Variational Gaussian Process Classifiers,” 1997.
- H. Kim and Z. Ghahramani, “The EM-EP Algorithm for Gaussian Process Classification,” *Proceedings of the Workshop on Probabilistic Graphical Models for Classification* (at ECML), 2003
- M. Kuss and C. Rasmussen, “Assessing Approximations for Gaussian Process Classification,” NIPS 2005.
- D. MacKay, “Gaussian Processes: A Replacement for Supervised Neural Networks?” Lecture notes for a tutorial at NIPS 1997.
- R. Neal, “Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification,” Technical Report No. 9702, Department of Statistics, University of Toronto, 1997.
- C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
- C. Rasmussen, “Gaussian Processes in Machine Learning,” Advanced Lectures on Machine Learning: ML Summer Schools, Canberra, Australia, 2003.
- C. Williams and D. Barber, “Bayesian Classification with Gaussian Processes,” *IEEE PAMI* 20(12) pp. 1342-1351, 1998.