Introduction to Gaussian Process Regression

Hanna M. Wallach

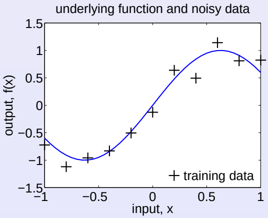
hmw26@cam.ac.uk

January 25, 2005

Outline

- Regression: weight-space view
- Regression: function-space view (Gaussian processes)
- Weight-space and function-space correspondence
- Making predictions
- Model selection: hyperparameters

Supervised Learning: Regression (1)



- Assume an underlying process which generates “clean” data.
- Goal: recover underlying process from noisy observed data.

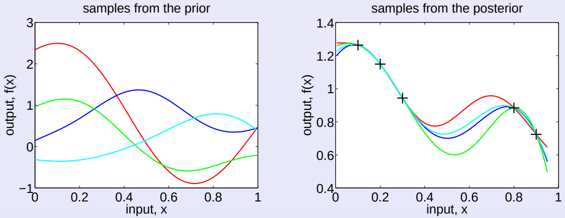
Supervised Learning: Regression (2)

- Training data are $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)} \mid i = 1, \ldots, n\}$.
- Each input is a vector $\mathbf{x}$ of dimension d.
- Each target is a real-valued scalar $y = f(\mathbf{x}) + \text{noise}$.
- Collect inputs in $d \times n$ matrix, X, and targets in vector, y:

$$\mathcal{D} = \{X, \mathbf{y}\}.$$

- Wish to infer $f^{\star}$ for unseen input $\mathbf{x}^{\star}$, using $P(f^{\star}|\mathbf{x}^{\star}, \mathcal{D})$.

Gaussian Process Models: Inference in Function Space

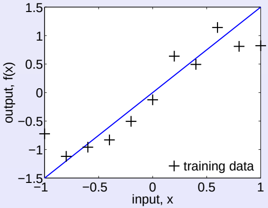


- A Gaussian process defines a distribution over functions.
- Inference takes place directly in function space.

Part I

Regression: The Weight-Space View

Bayesian Linear Regression (1)



- Assuming noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, the linear regression model is:

$$f(\mathbf{x}|\mathbf{w}) = \mathbf{x}^\top \mathbf{w}, \quad y = f + \epsilon.$$

Bayesian Linear Regression (2)

- Likelihood of parameters is:

$$P(\mathbf{y}|X, \mathbf{w}) = \mathcal{N}(X^\top \mathbf{w}, \sigma^2 I).$$

- Assume a Gaussian prior over parameters:

$$P(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \Sigma_p).$$

- Apply Bayes' theorem to obtain posterior:

$$P(\mathbf{w}|\mathbf{y}, X) \propto P(\mathbf{y}|X, \mathbf{w})P(\mathbf{w}).$$

Bayesian Linear Regression (3)

- Posterior distribution over **w** is:

$$P(\mathbf{w}|\mathbf{y}, X) = \mathcal{N}(\frac{1}{\sigma^2}A^{-1}X\mathbf{y}, A^{-1}) \text{ where } A = \Sigma_p^{-1} + \frac{1}{\sigma^2}XX^\top.$$

- Predictive distribution is:

$$\begin{aligned} P(f^\star|\mathbf{x}^\star, X, \mathbf{y}) &= \int f(\mathbf{x}^\star|\mathbf{w})P(\mathbf{w}|X, \mathbf{y})d\mathbf{w} \\ &= \mathcal{N}(\frac{1}{\sigma^2}\mathbf{x}^{\star\top}A^{-1}X\mathbf{y}, \mathbf{x}^{\star\top}A^{-1}\mathbf{x}^\star). \end{aligned}$$

Increasing Expressiveness

- Use a set of basis functions $\Phi(\mathbf{x})$ to project a d dimensional input $\mathbf{x}$ into m dimensional feature space:
  - *e.g.* $\Phi(x) = (1, x, x^2, \dots)$
- $P(f^\star|\mathbf{x}^\star, X, \mathbf{y})$ can be expressed in terms of inner products in feature space:
  - Can now use the kernel trick.
- How many basis functions should we use?

Part II

Regression: The Function-Space View

Gaussian Processes: Definition

- A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.
- Consistency:
  - If the GP specifies $y^{(1)}, y^{(2)} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, then it must also specify $y^{(1)} \sim \mathcal{N}(\mu_1, \Sigma_{11})$:
- A GP is completely specified by a mean function and a positive definite covariance function.

Gaussian Processes: A Distribution over Functions

- *e.g.* Choose mean function zero, and covariance function:

$$K_{p,q} = \mathrm{Cov}(f(\mathbf{x}^{(p)}), f(\mathbf{x}^{(q)})) = K(\mathbf{x}^{(p)}, \mathbf{x}^{(q)})$$

- For any set of inputs $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ we may compute K which defines a joint distribution over function values:

$$f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(n)}) \sim \mathcal{N}(\mathbf{0}, K).$$

- Therefore a GP specifies a distribution over functions.

Gaussian Processes: Simple Example

- Can obtain a GP from the Bayesin linear regression model:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \text{ with } \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p).$$

- Mean function is given by:

$$\mathbb{E}[f(\mathbf{x})] = \mathbf{x}^\top \mathbb{E}[\mathbf{w}] = 0.$$

- Covariance function is given by:

$$\mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] = \mathbf{x}^\top \mathbb{E}[\mathbf{w}\mathbf{w}^\top]\mathbf{x}' = \mathbf{x}^\top \Sigma_p \mathbf{x}'.$$

Weight-Space and Function Space Correspondence

- For any set of m basis functions, $\Phi(\mathbf{x})$, the corresponding covariance function is:

$$K(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) = \Phi(\mathbf{x}^{(p)})^\top \Sigma_p \Phi(\mathbf{x}^{(q)}).$$

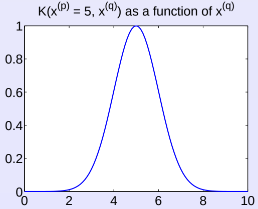
- Conversely, for every covariance function k, there is a possibly infinite expansion in terms of basis functions:

$$K(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) = \sum_{i=1}^{\infty} \lambda_i \Phi_i(\mathbf{x}^{(p)}) \Phi_i(\mathbf{x}^{(q)}).$$

The Covariance Function

- Specifies the covariance between pairs of random variables.
- *e.g.* Squared exponential covariance function:

$$\mathrm{Cov}(f(\mathbf{x}^{(p)}), f(\mathbf{x}^{(q)})) = K(\mathbf{x}^{(p)}, \mathbf{x}^{(q)}) = \exp\left(-\frac{1}{2}|\mathbf{x}^{(p)} - \mathbf{x}^{(q)}|^2\right).$$

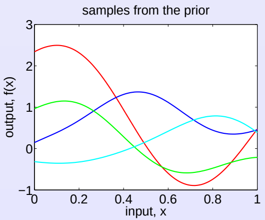


Gaussian Process Prior

- Given a set of inputs $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ we may draw samples $f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(n)})$ from the GP prior:

$$f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(n)}) \sim \mathcal{N}(\mathbf{0}, K).$$

- Four samples:



Posterior: Noise-Free Observations (1)

- Given noise-free training data:

$$\mathcal{D} = \{\mathbf{x}^{(i)}, f^{(i)} \mid i = 1, \ldots, n\} = \{X, \mathbf{f}\}.$$

- Want to make predictions $\mathbf{f}^\star$ at test points $X^\star$.
- According to GP prior, joint distribution of $\mathbf{f}$ and $\mathbf{f}^\star$ is:

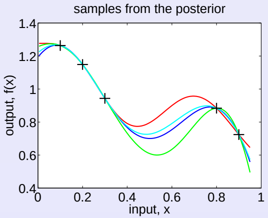
$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^\star \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X, X) & K(X, X^\star) \\ K(X^\star, X) & K(X^\star, X^\star) \end{bmatrix}\right).$$

Posterior: Noise-Free Observations (2)

- Condition $\{X^\star, \mathbf{f}^\star\}$ on $D = \{X, \mathbf{f}\}$ obtain the posterior.
- Restrict prior to contain only functions which agree with D.
- The posterior, $P(\mathbf{f}^\star | X^\star, X, \mathbf{f})$, is Gaussian with:

$$\boldsymbol{\mu} = K(X, X^\star) K(X, X)^{-1} \mathbf{f}, \text{ and}$$
$$\Sigma = K(X^\star, X^\star) - K(X, X^\star) K(X, X)^{-1} K(X^\star, X).$$

Posterior: Noise-Free Observations (3)



- Samples all agree with the observations $D = \{X, \mathbf{f}\}$.
- Greatest variance is in regions with few training points.

Prediction: Noisy Observations

- Typically we have noisy observations:

$$\mathcal{D} = \{X, \mathbf{y}\}, \text{ where } \mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon}$$

- Assume additive noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$.
- Conditioning on $D = \{X, \mathbf{y}\}$ gives a Gaussian with:

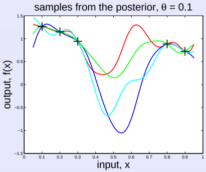
$$\boldsymbol{\mu} = K(X, X^\star)[K(X, X) + \sigma^2 I]^{-1}\mathbf{y}, \text{ and}$$
$$\Sigma = K(X^\star, X^\star) - K(X, X^\star)[K(X, X) + \sigma^2 I]^{-1} K(X^\star, X).$$

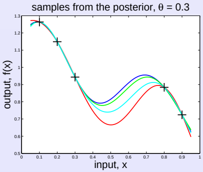
Model Selection: Hyperparameters

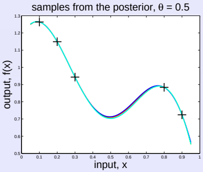
- *e.g.* the ARD covariance function:

$$k(x^{(p)}, x^{(q)}) = \exp\left(-\frac{1}{2\theta^2}(x^{(p)} - x^{(q)})^2\right).$$

- How best to choose θ?







Model Selection: Optimizing Marginal Likelihood (1)

- In absence of a strong prior $P(\theta)$, the posterior for hyperparameter θ is proportional to the marginal likelihood:

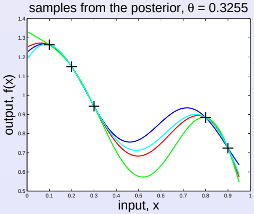
$$P(\theta|X, \mathbf{y}) \propto P(\mathbf{y}|X, \theta)$$

- Choose θ to optimize the marginal log-likelihood:

$$\log P(\mathbf{y}|X, \theta) = -\frac{1}{2} \log |K(X, X) + \sigma^2 I| - \frac{1}{2}\mathbf{y}^\top (K(X, X) + \sigma^2 I)^{-1} \mathbf{y} - \frac{n}{2} \log 2\pi.$$

Model Selection: Optimizing Marginal Likelihood (2)

- $\theta^{ML} = 0.3255$:



- Using θ^{ML} is an approximation to the true Bayesian method of integrating over all θ values weighted by their posterior.

References

1 Carl Edward Rasmussen. Gaussian Processes in Machine learning. *Machine Learning Summer School*, T ubingen, 2003. `http://www.kyb.tuebingen.mpg.de/~carl/mlss03/`

2 Carl Edward Rasmussen and Chris Williams. Gaussian Processes for Machine Learning. Forthcoming.

3 Carl Edward Rasmussen. The Gaussian Process Website. `http://www.gatsby.ucl.ac.uk/~edward/gp/`