732A96/TDDE15 Advanced Machine Learning

Gaussian Process Regression and Classification

Jose M. Peña
IDA, Linköping University, Sweden

Lectures 10: Gaussian Process Regression

Contents

- Linear Regression
- Bayesian Linear Regression
- Gaussian Process Regression
- Squared Exponential Covariance Function
- Gaussian Process Regression: Canadian Wages

Literature

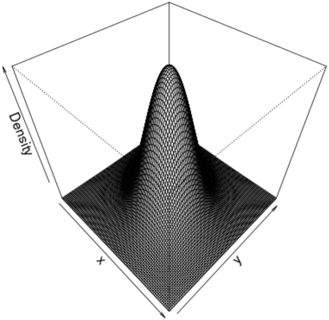
- Main source
  - Rasmussen, C. E. and Williams, K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006. Chapters 2.1-2.5.
- Additional source
  - Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. Chapters 6.4.1-6.4.2.

Gaussian Distribution

- Density function of the Gaussian (a.k.a normal) distribution for a D-dimensional random variable $\boldsymbol{X}$:

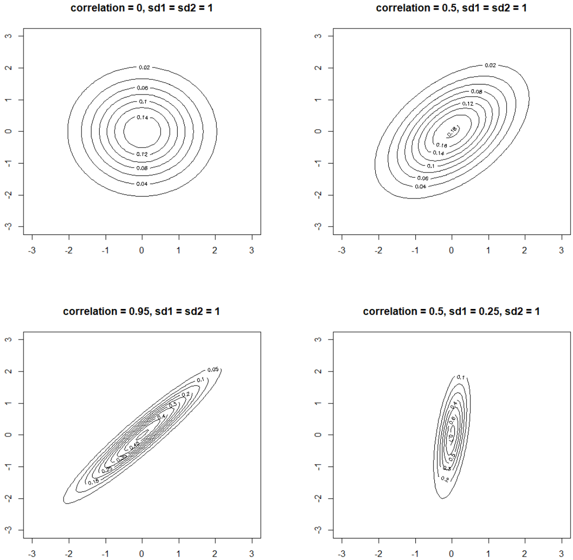
$$\mathcal{N}(\boldsymbol{x};\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu)\right\}$$

- Recall that $E[\boldsymbol{X}]=\mu$ and $cov(\boldsymbol{X})=\Sigma$.



Gaussian Distribution

- Example: $\mathcal{N}(x_1, x_2; \mu, \Sigma)$ with $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$.



correlation = 0, sd1 = sd2 = 1

correlation = 0.5, sd1 = sd2 = 1

correlation = 0.95, sd1 = sd2 = 1

correlation = 0.5, sd1 = 0.25, sd2 = 1

Gaussian Distribution

- Recall that if

$$p(x) = \mathcal{N}(x; \mu, \Lambda^{-1})$$
$$p(y|x) = \mathcal{N}(y; Ax + B, L^{-1})$$

then

$$p(x, y) = \mathcal{N}(x, y; A\mu + B, R^{-1})$$

where

$$R^{-1} = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}.$$

- Recall also that if $p(x) = \mathcal{N}(x; \mu, \Sigma)$ and $\Lambda = \Sigma^{-1}$ and

$$x = (x_a, x_b)^T \qquad \mu = (\mu_a, \mu_b)^T$$
$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \qquad \Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

then

$$p(x_a) = \mathcal{N}(x_a; \mu_a, \Sigma_{aa})$$
$$p(x_a|x_b) = \mathcal{N}(x_a; \mu_{a|b}, \Lambda_{aa}^{-1}) \qquad \text{where } \mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b) \text{ or}$$
$$p(x_a|x_b) = \mathcal{N}(x_a; \mu_{a|b}, \Sigma_{a|b}) \qquad \text{where } \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$
$$\text{and } \Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}.$$

Linear Regression

- Training data: $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) | i = 1, \ldots, n\} = (X, \boldsymbol{y})$.
- Deterministic function: $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w}$.
- Additive noisy observations: $y = f(\boldsymbol{x}) + \epsilon$.
- Gaussian noise: $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.
- Likelihood function: $p(\boldsymbol{y}|X, \boldsymbol{w}) = \mathcal{N}(X^T \boldsymbol{w}, \sigma_n^2 I) \propto \exp\left\{\frac{1}{2\sigma_n^2} ||\boldsymbol{y} - X^T \boldsymbol{w}||^2\right\}$.
- To obtain $\boldsymbol{w}^{ML}$,
  - take the derivative of the log lik function wrt $\boldsymbol{w}$, and
  - set it to zero, and
  - solve to obtain $\boldsymbol{w}^{ML} = (XX^T)^{-1} X \boldsymbol{y}$.
- Minimizing the least squared error (i.e., $\frac{1}{2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{w})^2$) gives the same result. This justifies the use of LSE.

Bayesian Linear Regression

- Prior distribution: $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma_p)$, e.g. ridge regression $\Sigma_p = \alpha^{-1} I$.
- Posterior distribution:

$$\log p(\boldsymbol{w}|X, \boldsymbol{y}) \propto \log p(\boldsymbol{y}|X, \boldsymbol{w}) + \log p(\boldsymbol{w}) \propto \frac{1}{2\sigma_n^2} \|\boldsymbol{y} - X^T \boldsymbol{w}\|^2 - \frac{1}{2} \boldsymbol{w}^T \Sigma_p^{-1} \boldsymbol{w}.$$

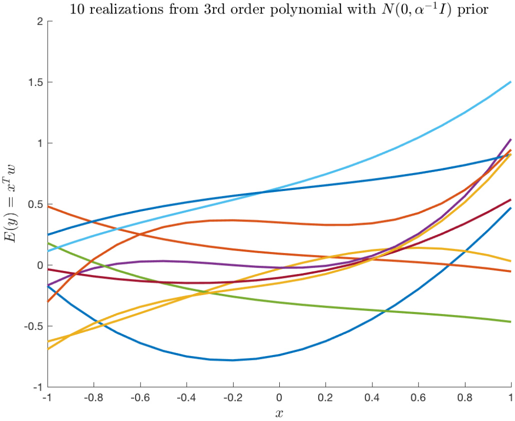
- So, $\boldsymbol{w}^{MAP}$ can be seen as a penalized/regularized ML estimate.
- Specifically, $p(\boldsymbol{w}|X, \boldsymbol{y}) = \mathcal{N}(\bar{\boldsymbol{w}} = \frac{1}{\sigma_n^2} A^{-1} X \boldsymbol{y}, A^{-1})$ where $A = \sigma_n^{-2} X X^T + \Sigma_p^{-1}$, and thus $\boldsymbol{w}^{MAP} = \bar{\boldsymbol{w}}$.
- A full Bayesian approach does not use $\boldsymbol{w}^{MAP}$ but the predictive distribution:

$$p(f_*|\boldsymbol{x}_*, X, \boldsymbol{y}) = \int p(f_*|\boldsymbol{x}_*, \boldsymbol{w}) p(\boldsymbol{w}|X, \boldsymbol{y}) d\boldsymbol{w} = \mathcal{N}(\frac{1}{\sigma_n^2} \boldsymbol{x}_* A^{-1} X \boldsymbol{y}, \boldsymbol{x}_*^T A^{-1} \boldsymbol{x}_*).$$

- The above carries over to the feature space $\phi(\boldsymbol{x})$. The kernel trick applies. See Section 2.1.2 in Rasmussen and Williams.

Bayesian Linear Regression

- A prior on $\boldsymbol{w}$ is a prior on f.

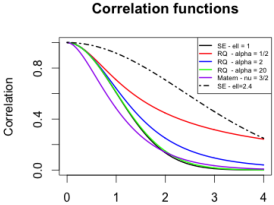


Gaussian Processes Regression

- A GP defines a prior distribution **over functions directly**, instead of indirectly through weights as before. Therefore, a GP operates on the space of functions rather than on the space of weights. Operating in either space is equivalent. A GP defines a prior over functions by defining a prior over a **finite** number of input points.
- Formally, a GP is a collection of random variables, any finite number of which have a joint Gaussian distribution. Hence, a GP is defined as
  - $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ where
  - $m(\mathbf{x}) = E[f(\mathbf{x})]$ is the mean function (assumed to be zero hereinafter), and
  - $k(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$ is the covariance function, e.g. squared exponential:

$$k(\mathbf{x}, \mathbf{x}') = cov(f(\mathbf{x}), f(\mathbf{x}')) = \sigma_f^2 \exp\left\{ - \frac{||\mathbf{x} - \mathbf{x}'||^2}{2\ell^2} \right\}$$

i.e. highly correlated **function values** for close **input points**. Intuitively, σ_f^2 is the overall variance of the function, and ℓ is the distance we have to move in the input space for the function to vary significantly.



Gaussian Processes Regression

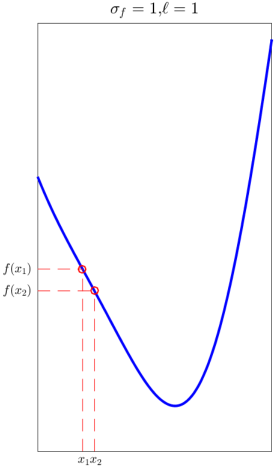
- Formally, a GP is a collection of random variables, any finite number of which have a joint Gaussian distribution. Hence, a GP is defined as
  - $f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$ where
  - $m(\boldsymbol{x}) = E[f(\boldsymbol{x})]$ is the mean function (assumed to be zero hereinafter), and
  - $k(\boldsymbol{x}, \boldsymbol{x}') = E[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))]$ is the covariance function, e.g. squared exponential:

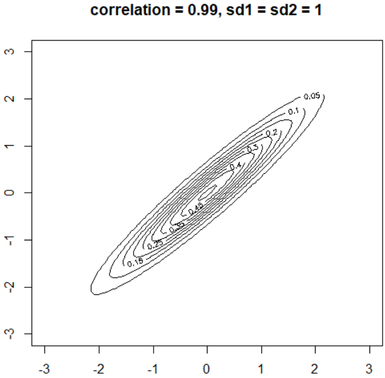
$$k(\boldsymbol{x}, \boldsymbol{x}') = cov(f(\boldsymbol{x}), f(\boldsymbol{x}')) = \sigma_f^2 \exp\Big\{ - \frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{2\ell^2} \Big\}.$$

- Note that each random variable or dimension in a GP is a **function value** at an input point. Hence, a GP specifies a probability distribution over functions at any **finite** number of input points.
- Note that the covariance between the **function values** is written as a function of the **inputs points**. Note that the covariances are non-negative.
- Note that the zero mean assumption affects the location but not the shape of the prior. Note also that the posterior mean is not assumed to be zero. However, it may get close to zero in regions with few input points when ℓ is small, because the posterior is then similar to the prior. Of course, one can always center or standarize the data.
- We can sample the function space by sampling the GP at any number of chosen input points X_*. To do so, we sample a multivariate Gaussian distribution with the corresponding covariance matrix, i.e. $\boldsymbol{f}_* | X_* \sim \mathcal{N}(0, K(X_*, X_*))$.
- Demo of `GaussianProcesses.R`.

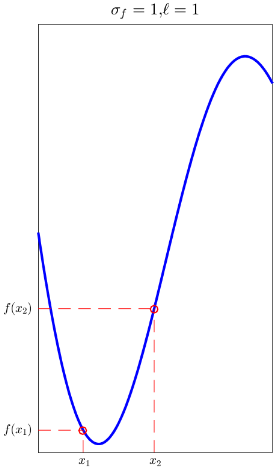
Squared Exponential Covariance: Smooth Function, Close Points

- If $\sigma_f = 1$, then $k(x,x) = 1$, $0 \leq k(x,x') \leq 1$, and $k(x,x') = \rho(f(x), f(x'))$.

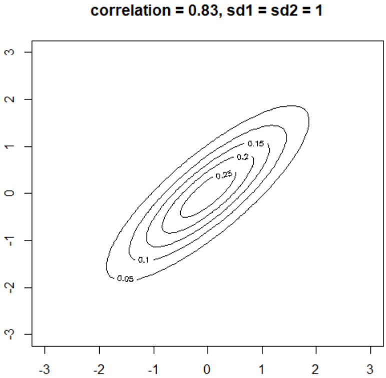




Squared Exponential Covariance: Smooth Function, Distant Points



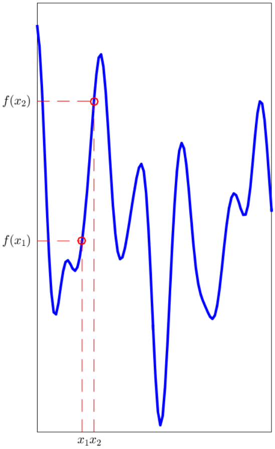
$\sigma_f = 1, \ell = 1$



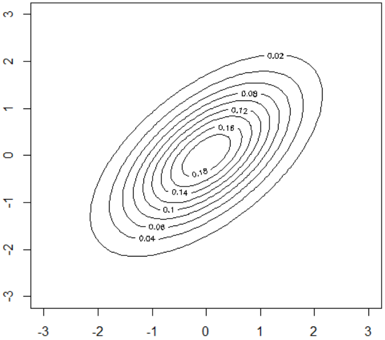
correlation = 0.83, sd1 = sd2 = 1

Squared Exponential Covariance: Jagged Function, Close Points

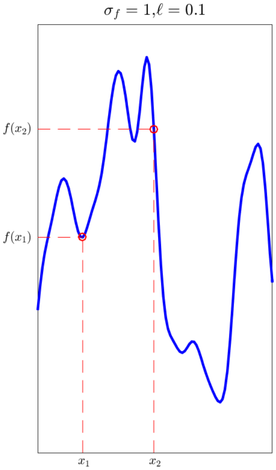
$\sigma_f = 1, \ell = 0.1$

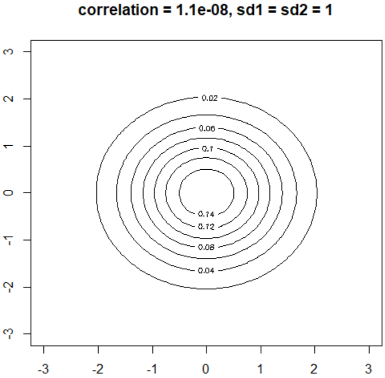


correlation = 0.6, sd1 = sd2 = 1



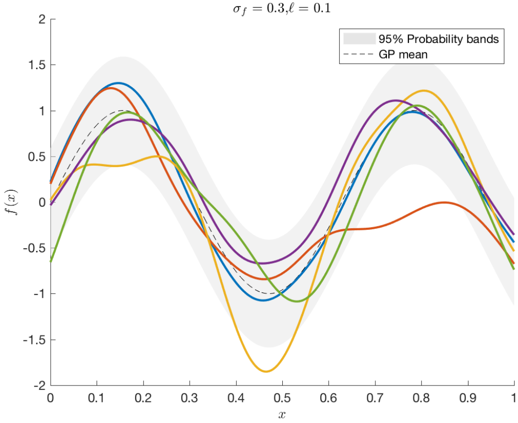
Squared Exponential Covariance: Jagged Function, Distant Points





Gaussian Process Sampling: Multivariate Draw

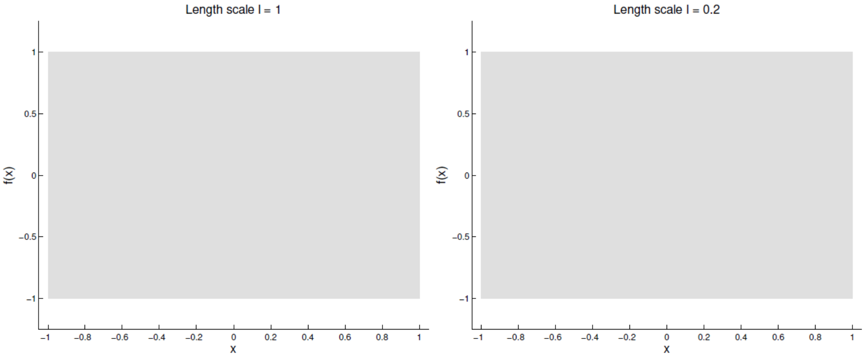
- To sample a GP at points $X_* = \{x_1, \ldots, x_n\}$, we sample a multivariate Gaussian distribution $\mathcal{N}(0, K(X_*, X_*))$.



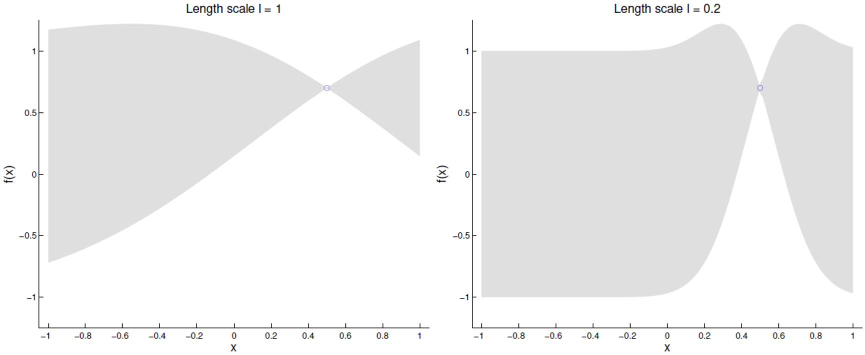
Gaussian Process Sampling: Before the First Univariate Draw

- To sample a GP at points $X_* = \{x_1, \ldots, x_n\}$, we can alternatively sample univariate Gaussian distributions, since

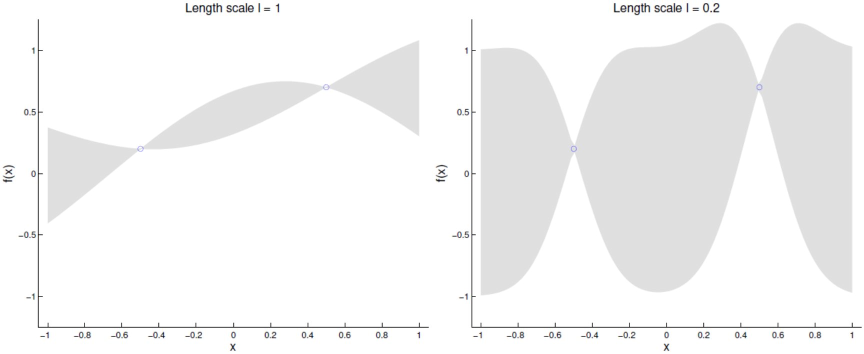
$$p(f(x_1), \ldots, f(x_n)) = \prod_{i=1}^{n} p(f(x_i) | f(x_1), \ldots, f(x_{i-1})).$$



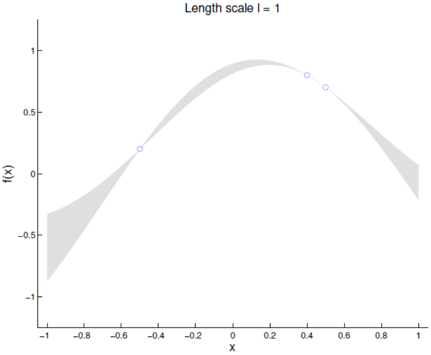
Gaussian Process Sampling: Before the Second Univariate Draw

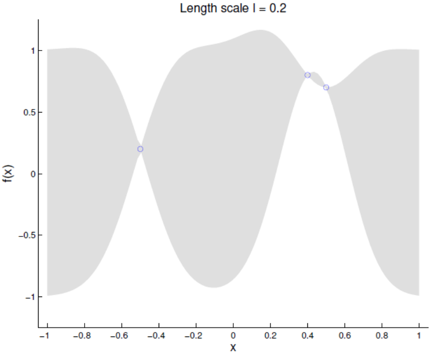


Gaussian Process Sampling: Before the Third Univariate Draw



Gaussian Process Sampling: Before the Fourth Univariate Draw





Gaussian Processes Regression

- With no training data, sample from $\boldsymbol{f}_*|X_* \sim \mathcal{N}(0, K(X_*, X_*))$.
- With **noise-free** training data $\mathcal{D} = \{(\boldsymbol{x}_i, f_i)|i = 1, \ldots, n\} = (X, \boldsymbol{f})$, build

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

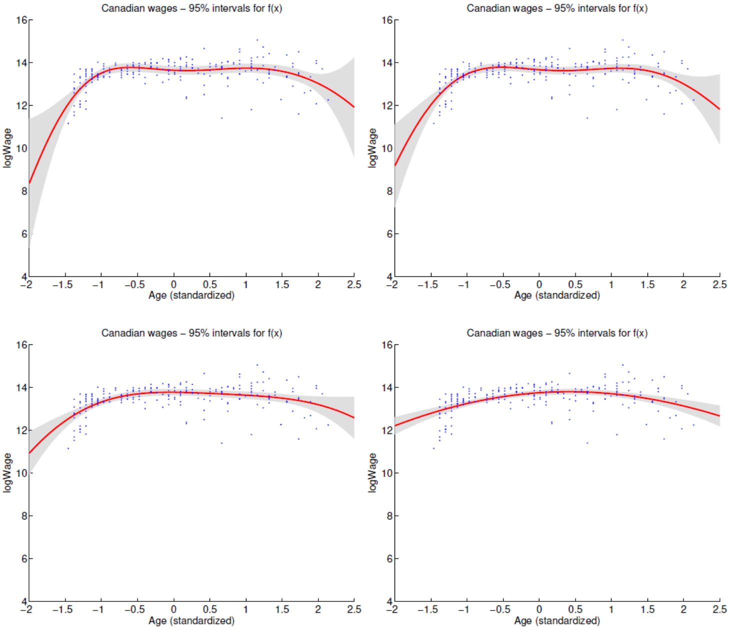
  and sample from $\boldsymbol{f}_*|X_*, X, \boldsymbol{f} \sim \mathcal{N}(K(X_*,X)K(X,X)^{-1}\boldsymbol{f}, K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*))$.
- With **noisy** training data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)|i = 1, \ldots, n\} = (X, \boldsymbol{y})$, build[1]

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

  and sample from $\boldsymbol{f}_*|X_*, X, \boldsymbol{y} \sim \mathcal{N}(K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1}\boldsymbol{y}, K(X_*,X_*) - K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1}K(X,X_*))$.
- See p. 17 of Rasmussen and Williams for the correspondence between the weight and function space views: Every covariance function can be mapped into a set of features, and vice versa.
- Demo of `KernLabDemo.R`.

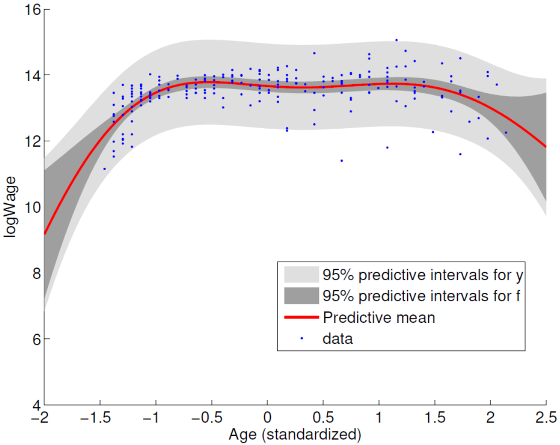
[1] $cov(y, y') = cov(f(x) + \epsilon, f(x') + \epsilon') = cov(f(x), f(x')) + cov(f(x), \epsilon') + cov(\epsilon, f(x')) + cov(\epsilon, \epsilon') = cov(f(x), f(x')) + \sigma_n^2 \delta_{xx'}$.

Gaussian Process Regression: Canadian Wages ($\ell = 0.2, 0.5, 1, 2$)



Gaussian Process Regression: Canadian Wages ($\ell = 0.5$)

- Predictive interval for $\boldsymbol{f}_*$: mean($\boldsymbol{f}_*$) $\pm$ 1.96 sqrt(var($\boldsymbol{f}_*$)).
- Predictive interval for $\boldsymbol{y}_*$: mean($\boldsymbol{f}_*$) $\pm$ 1.96 sqrt(var($\boldsymbol{f}_*$) $+ \sigma_n^2$).



Contents

- Linear Regression
- Bayesian Linear Regression
- Gaussian Processes Regression
- Squared Exponential Covariance Function
- Gaussian Process Regression: Canadian Wages

Thank you