Procesos Gaussianos

Eduardo Morales

INAOE

Contenido

Procesos Gaussianos

- Normalmente, en los algoritmos de aprendizaje que hemos visto, dado un conjunto de ejemplos de entrenamiento se busca encontrar el mejor modelo que ajuste a los datos (o el modelo que haga las mejores predicciones con ejemplos de prueba).
- Cuando hablamos de algoritmos bayesianos, en cambio podemos buscar la distribución posterior sobre los modelos.
- Estas distribuciones nos cuantifican nuestra incertidumbre en los modelos.

Procesos Gaussianos

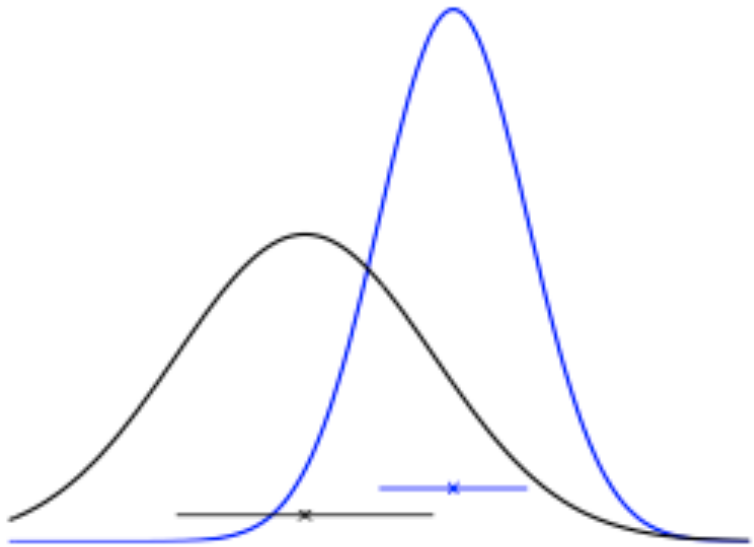
- Los procesos gaussianos se definen como una distribución de probabilidad sobre *funciones* aleatorias.
- De hecho son sobre colecciones infinitas de variables (funciones), tal que cualquier subconjunto de variables aleatoria finita tiene una distribución gaussiana multivariable.

Procesos Gaussianos

- Los procesos gaussianos se empezaron a estudiar para predicción de series de tiempo en los 40's (Wiener, Kolmogorov)
- Más adelante (70's) se usaron en geoestadística y meteorología, en lo que se llamó *kriging* usando 2 o 3 dimensiones.
- Más adelante en estadística espacial (ver Creesie, 1993)
- También en regresión: O'Hagan, 1978
- Se realizaron experimentos en computadoras (sin ruido) a finales de los 80's (Sacks et al. 1989)
- Se popularizaron recientemente en aprendizaje computacional por Williams y Rasmussen (1996) y Neal (1996) para resolver problemas de regresión.

Distribución Gaussiana

La distribución Gaussiana es la más conocida y utilizada en probabilidad:



Distribución Gaussiana

$$\begin{aligned} p(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2}) \\ &= \mathcal{N}(y|\mu,\sigma^2) \end{aligned}$$

- El valor esperado es: $E\{x\} = \mu$
- La varianza es: $Var(x) = \sigma^2$
- La desviación estandar es: σ

Propiedades

- La suma de gaussianas también es una gaussiana:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$\sum_{i=1}^{n} y_i \sim \mathcal{N}(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$$

- Conforme aumenta la sumatoria, la suma de variables aleatorias independientes (no necesariamente gaussianas con varianza finita) tienden a una gaussiana (Teorema del Límite Central)

Propiedades

- Si escalamos una gaussiana, también es gaussiana:

$$y_i \sim \mathcal{N}(\mu, \sigma^2)$$

$$wy \sim \mathcal{N}(w\mu, w^2\sigma^2)$$

- La distribución gaussiana se puede extender hacia gaussianas multivariables.
- Ahora se tiene un vector media $\mu \in \mathbf{R}^n$ y matriz de covarianza Σ de $n \times n$ que es simétrica definida positiva ($x^T Ax > 0$ y todos su eigenvalores son positivos).

Gaussianas Multivariables

- Las distribuciones gaussianas multivariables son útiles para modelar colecciones finitas de variables contínuas.

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

Gaussianas Multivariables

Ejemplo de dos variables:

$$p(w,h) = \frac{1}{\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2}} exp(-\frac{1}{2}(\frac{(w-\mu_1)^2}{\sigma_1^2} + \frac{(h-\mu_2)^2}{\sigma_2^2}))$$

$$p(w,h) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} exp(-\frac{1}{2}\left(\left(\begin{bmatrix} w \\ h \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^T \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \left(\begin{bmatrix} w \\ h \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)\right)$$

Distribución Gaussiana Multivariable

- En general, la definición de una distribución gaussiana multivariable es:

$$p(X|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu))$$

donde X es un vector de variables y Σ es una matriz de convarianza entre las variables

Procesos Gaussianos

- Los procesos gaussianos son una extensión hacia colecciones infinitas de variables.
- Esta extensión nos permite pensar en los procesos gaussianos como distribuciones, no sólo sobre vectores aleatorios sino sobre distribuciones de funciones aleatorias.
- Para entender esto, pongamos un ejemplo sencillo

Procesos Gaussianos

- Vamos a suponer que tenemos un conjunto de posibles funciones que mapean un vector X a $\mathcal{R}$. Algunos ejemplos de una función particular h_0 podría ser: $h_0(x_1) = 5$, $h_0(x_2) = 2.3$, $h_0(x_3) = \pi$, ..., $h_0(x_m) = -7$
- Como el dominio de h tiene solo m elementos, podemos representarlo de forma compacta como un vector $\vec{h} = [h(x_1), h(x_2), \ldots, h(x_m)]^T$
- Para especificar una distribución de probabilidad, necesitamos asociarle una densidad de probabilidad a cada posible función h.

Procesos Gaussianos

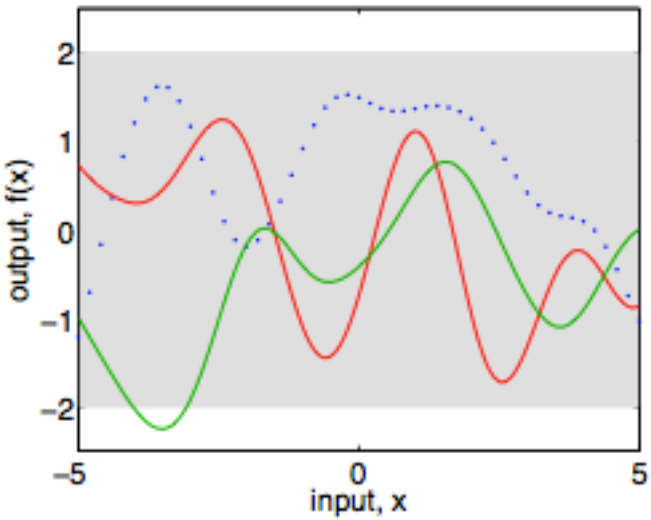
- Una forma natural es hacer una correspondencia entre la función y su vector $\vec{h}$, $\vec{h} = \mathcal{N}(\vec{\mu}, \sigma^2 I)$ (y suponiendo i.i.d):

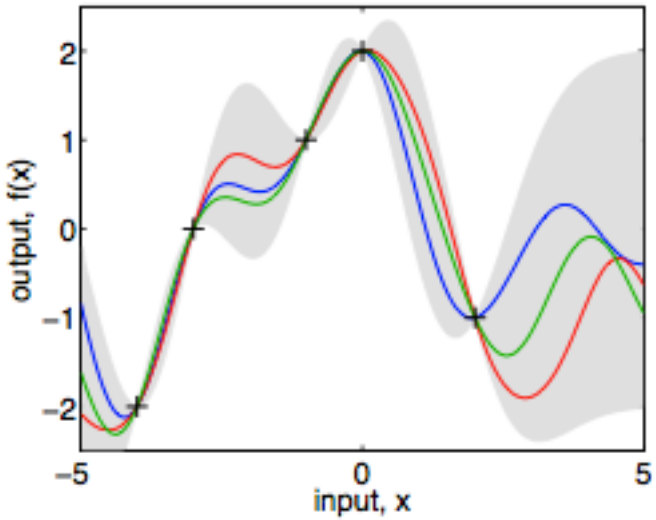
$$P(h) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}(h(x_i) - \mu_i)^2)$$

- Esto nos permite asociar distribuciones de probabilidades sobre funciones de dominios finitos usando una distribución multivariable gaussiana finita, sobre funciones de salida $h(x_1), \ldots, h(x_m)$ en un número finito de puntos de entrada $x_1, \ldots, x_m$

Procesos Gaussianos

- ¿Cómo podemos hacerlo cuando el dominio es infinito?





Procesos Gaussianos

- Un proceso gaussiano es un proceso estocástico tal que cualquier subconjunto de variables aleatoria finita tiene una distribución gaussiana multivariable.
- En particular, una colección de variables aleatorias $\{h(x) : x \in \mathcal{X}\}$ se obtiene de un proceso gaussiano con una función media $m(\cdot)$ y una función de covarianza $k(\cdot, \cdot)$, si para cualquier conjunto finito de elementos $x_1, \ldots, x_m \in \mathcal{X}$, el conjunto finito de variables aleatorias asociadas $h(x_1), \ldots, h(x_m)$ tienen la siguiente ditribución

Procesos Gaussianos

$$\begin{bmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_m) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & \dots & k(x_m, x_m) \end{bmatrix} \right)$$

Esto lo denotamos como: $h(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot))$

Procesos Gaussianos

- Cada dimensión de la gaussiana corresponde a un elemento x y su componente correspondiente del vector aleatorio representa el de $h(x)$.
- Usando las propiedades de marginalización para gaussianas multivariables, podemos obtener la densidad marginal multivariable gaussiana correspondiente a cualquier subconjunto finito de variables.
- Para $m(\cdot)$ podemos usar cualquier función real, pero para $k(\cdot,\cdot)$ debe de cumplirse que para cualquier conjunto de elementos $x_1, \ldots, x_m \in \mathcal{X}$, la matriz resultante:

Procesos Gaussianos

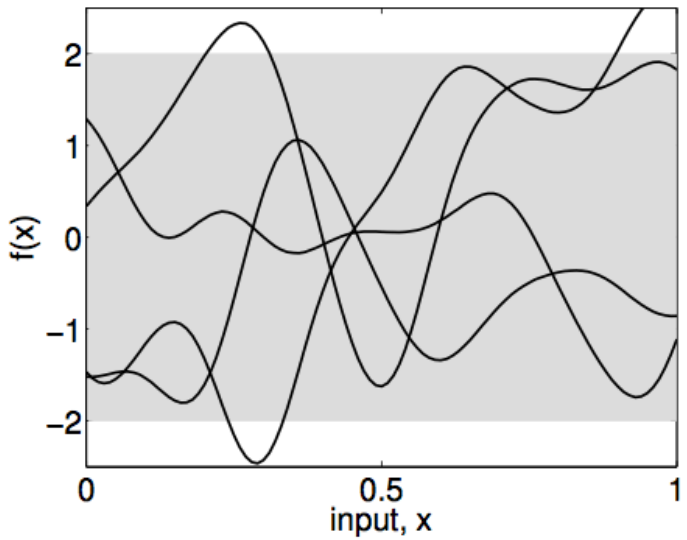
$$K = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & \dots & k(x_m, x_m) \end{bmatrix}$$

sea válida para una distribución gaussiana multivariable (positiva semidefinitiva), lo cual son las mismas condiciones que para los Kernels (Mercer's condition), por lo que cualquier función kernel se puede usar como función de covarianza.

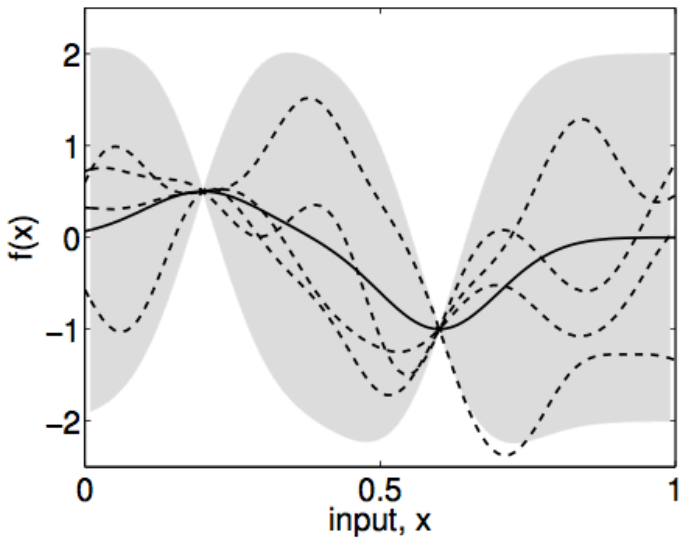
Procesos Gaussianos

- Uno de los esquemas más usados en GPs es considerar una media cero y usar el kernel exponencial cuadrado:
  $h(\cdot) \sim \mathcal{GP}(0, k(\cdot,\cdot))$
  $k_{SE}(x, x') = exp(-\frac{1}{2\tau^2}||x - x'||^2)$
- con este kernel:
  - $h(x)$ y $h(x')$ tienen alta covarianza cuando x y x' están cercanos en el espacio de entrada, i.e.,
    $||x - x'|| = |x - x'| \approx 0$ y $exp(-\frac{1}{2\tau^2}||x - x'||^2) \approx 1$
  - $h(x)$ y $h(x')$ tienen baja covarianza cuando x y x' están separados en el espacio de entrada, i.e., $||x - x'|| \gg 0$ y $exp(-\frac{1}{2\tau^2}||x - x'||^2) \approx 0$

Procesos Gaussianos



(a), prior



(b), posterior

Procesos Gaussianos

- El modelo de regresión de un proceso gaussiano: Sea $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$ el conjunto de entrenamiento de ejemplos i.i.d. tomados de una distribución desconocida.
- En un modelo de regresión gaussiano:

$$y^{(i)} = h(x^{(i)}) + \epsilon^{(i)}, \text{ con } i = 1, \ldots, m$$

donde $\epsilon^{(i)}$ son variables de ruido i.i.d. con distribuciones $\mathcal{N}(0, \sigma^2)$ independientes.

Procesos Gaussianos

- Suponemos una distribución *a priori* sobre las funciones $h(\cdot)$, en particular, una gaussiana con media zero: $h(\cdot) \sim \mathcal{GP}(0, k(\cdot,\cdot))$
- Sean $T = \{(x_*^{(i)}, y_*^{(i)})\}_{i=1}^{m_*}$ un conjunto de datos de pruebas i.i.d. tomados de la misma distribución.
- Para las regresión lineal bayesiana usamos la regla de bayes para calcular la distribución predictiva posterior, para los procesos gaussianos existe una solución "más sencilla"!

Procesos Gaussianos

- En particular, se debe cumplir que:

$$p\left(\begin{bmatrix} \vec{h} \\ \vec{h_*} \end{bmatrix} | X, X_* \right) \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} k(X,X) & k(X,X_*) \\ k(X_*,X) & k(X_*,X_*) \end{bmatrix}\right)$$

Procesos Gaussianos

donde:

$$\vec{h} \in \mathbf{R}^m \text{ tal que } \vec{h} = [h(x^{(1)}) \dots h(x^{(m)})]^T$$

$$\vec{h_*} \in \mathbf{R}^{m_*} \text{ tal que } \vec{h_*} = [h(x_*^{(1)}) \dots h(x_*^{(m)})]^T$$

$$K(X, X) \in \mathbf{R}^{m \times m} \text{ tal que } (K(X, X))_{i,j} = k(x^{(i)}, x^{(j)})$$

$$K(X, X_*) \in \mathbf{R}^{m \times m} \text{ tal que } (K(X, X))_{i,j} = k(x^{(i)}, x_*^{(j)})$$

$$K(X_*, X) \in \mathbf{R}^{m \times m} \text{ tal que } (K(X, X))_{i,j} = k(x_*^{(i)}, x^{(j)})$$

$$K(X_*, X_*) \in \mathbf{R}^{m \times m} \text{ tal que } (K(X, X))_{i,j} = k(x_*^{(i)}, x_*^{(j)})$$

Procesos Gaussianos

- Para el ruido:

$$p\left(\begin{bmatrix} \vec{\epsilon} \\ \vec{\epsilon_*} \end{bmatrix}\right) \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} \sigma^2 I & \vec{0} \\ \vec{0}^T & \sigma^2 I \end{bmatrix}\right)$$

Procesos Gaussianos

- Suponemos que son independientes, por lo que su suma también lo es:

$$\begin{bmatrix} \vec{y} \\ \vec{y_*} \end{bmatrix} | X, X_* = \begin{bmatrix} \vec{h} \\ \vec{h_*} \end{bmatrix} + \begin{bmatrix} \vec{\epsilon} \\ \vec{\epsilon_*} \end{bmatrix} \sim$$

$$\mathcal{N}\left(\vec{0}, \begin{bmatrix} k(X,X) + \sigma^2 I & k(X,X_*) \\ k(X_*,X) & k(X_*,X_*) + \sigma^2 I \end{bmatrix}\right)$$

Procesos Gaussianos

- Usando las reglas de condicionamiento gaussianas, se sigue que:

$$\vec{y}_*|\vec{y}, X, X_* \sim \mathcal{N}(\mu^*, \Sigma^*)$$

donde:

$$\mu^* = K(X_*, X)(K(X, X) + \sigma^2 I)^{-1}\vec{y}$$

$$\Sigma^* = K(X_*, X_*) + \sigma^2 I - K(X_*, X)(K(X, X) + \sigma^2 I)^{-1}K(X, X_*)$$

Gaussianas Multivariables

- Una propiedad importante de las distribuciones gaussianas, es que la probabilidad condicional de dos distribuciones gaussianas también es gaussiana.
- Vamos a suponer que tenemos dividido un conjunto de variables en dos subconjuntos:

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$$

y lo mismo para la media μ:

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

Gaussianas Multivariables

- y la matríz de covarianza Σ:

$$\Sigma = \begin{pmatrix} \Sigma_{a,a} & \Sigma_{a,b} \\ \Sigma_{b,a} & \Sigma_{b,b} \end{pmatrix}$$

- Se puede notar que $\Sigma^T = \Sigma$, que $\Sigma_{a,a}$ y $\Sigma_{b,b}$ son simétricas y que $\Sigma_{b,a} = \Sigma_{a,b}^T$

Gaussianas Multivariables

- Muchas veces es más conveniente trabajar con el inverso de la matriz de covarianza: $\Lambda \equiv \Sigma^{-1}$, también conocida como la matriz de precisión (*precision matrix*):

$$\Lambda = \begin{pmatrix} \Lambda_{a,a} & \Lambda_{a,b} \\ \Lambda_{b,a} & \Lambda_{b,b} \end{pmatrix}$$

  La cual también cumple con las condiciones de simetría.

Gaussianas Multivariables

- Ahora para calcular $p(x_a|x_b)$ se podría sacar de la probabilidad conjunta y normalizando, sin embargo, se puede obtener una solución más eficientemente.
- Si nos fijamos en el exponente de la distribución gaussiana:

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) =$$
$$-\frac{1}{2}(x_a-\mu_a)^T\Lambda_{a,a}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^T\Lambda_{a,b}(x_b-\mu_b)$$
$$-\frac{1}{2}(x_b-\mu_b)^T\Lambda_{b,a}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^T\Lambda_{b,b}(x_b-\mu_b)$$

Gaussianas Multivariables

- Nuestro objetivo es cambiar esta expresión por una que dependa de la media y varianza de $p(x_a|x_b)$.
- Esta es una operación común en distribuciones gaussianas, llamada "completando el cuadrado" que permita soluciones directas.
- En general el exponente de una gaussiana se puede expresar como:

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu + \text{ const}$$

donde *const* denota a términos que no dependen de x.

Gaussianas Multivariables

- En esta expresión, el término de segundo orden se asocia al inverso de la matriz de covarianza y el término lineal en x con $\Sigma^{-1}\mu$ de donde se puede obtener μ.
- Esto lo podemos encontrar por inspección en la expresión anterior, buscando todos los términos de segundo orden de x_a, en este caso, el único término es:

$$-\frac{1}{2}x_a^T \Lambda_{a,a} x_a$$

De donde podemos concluir que la covarianza (inverso de precisión) de $p(x_a|x_b)$ está dado por:

$$\Sigma_{a|b} = \Lambda_{a,a}^{-1}$$

Gaussianas Multivariables

- Ahora buscamos a todos los términos que sean lineales en x_a:

$$x_a^T\{\Lambda_{a,a}\mu_a - \Lambda_{a,b}(x_b - \mu_b)\}$$

donde aprovechamos que: $\Lambda_{b,a}^T = \Lambda_{a,b}$. Como esto debe de ser igual a: $\Sigma_{a|b}^{-1}\mu_{a|b}$, entonces (y usando $\Sigma_{a|b} = \Lambda_{a,a}^{-1}$):

$$\mu_{a|b} = \Sigma_{a|b}\{\Lambda_{a,a}\mu_a - \Lambda_{a,b}(x_b - \mu_b)\} = \mu_a - \Lambda_{a,a}^{-1}\Lambda_{a,b}(x_b - \mu_b)$$

Gaussianas Multivariables

- Para poder hacer uso de esto usamos una identididad de invesión de matrices partidas:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}$$

donde $M = (A - BD^{-1}C)^{-1}$. A M se le conoce como el *complemento Schur* de la matriz.

Gaussianas Multivariables

Con esta definición:

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{a,a} & \Sigma_{a,b} \\ \Sigma_{b,a} & \Sigma_{b,b} \end{pmatrix}^{-1} = \Lambda = \begin{pmatrix} \Lambda_{a,a} & \Lambda_{a,b} \\ \Lambda_{b,a} & \Lambda_{b,b} \end{pmatrix}$$

y

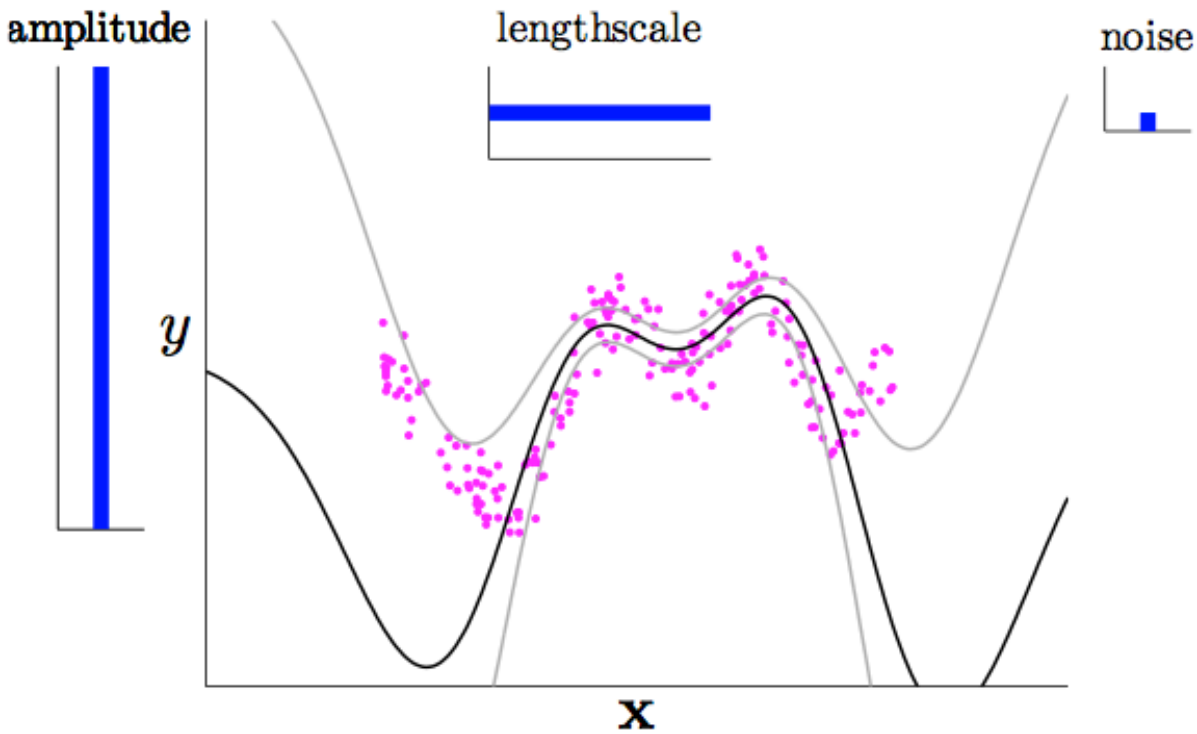
$$\Lambda_{a,a} = (\Sigma_{a,a} - \Sigma_{a,b}\Sigma_{b,b}^{-1}\Sigma_{b,a})^{-1}$$

$$\Lambda_{a,b} = -(\Sigma_{a,a} - \Sigma_{a,b}\Sigma_{b,b}^{-1}\Sigma_{b,a})^{-1}\Sigma_{a,b}\Sigma_{b,b}^{-1}$$

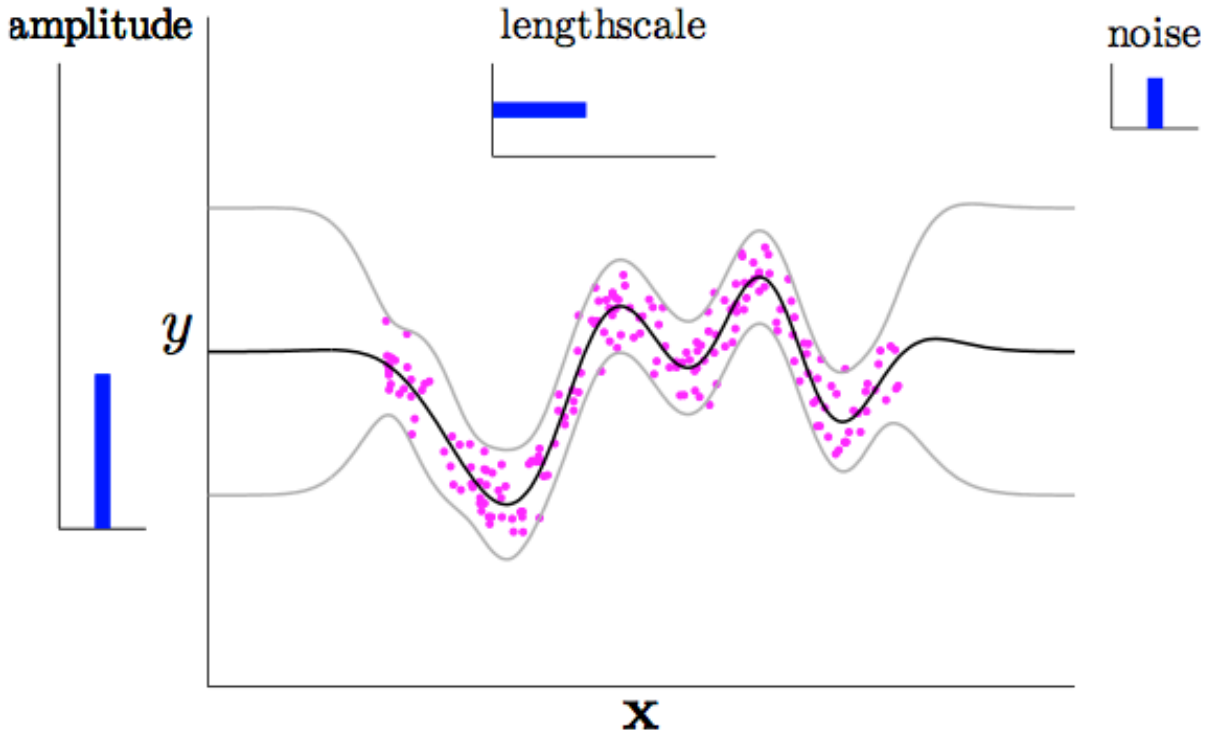
$$\mu_{a|b} = \mu_a - \Sigma_{a,b}\Sigma_{b,b}^{-1}(x_b - \mu_b)$$

$$\Sigma_{a|b} = \Sigma_{a,a} - \Sigma_{a,b}\Sigma_{b,b}^{-1}\Sigma_{b,a}$$

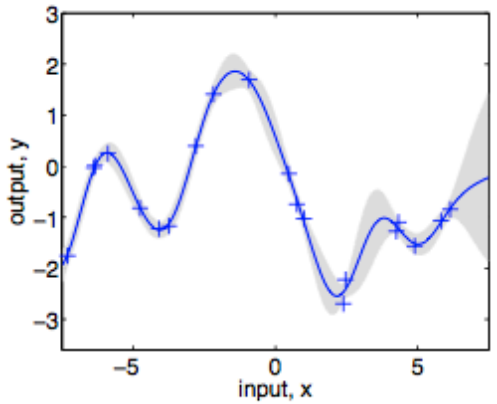
Efectos de los Hiperparámetros



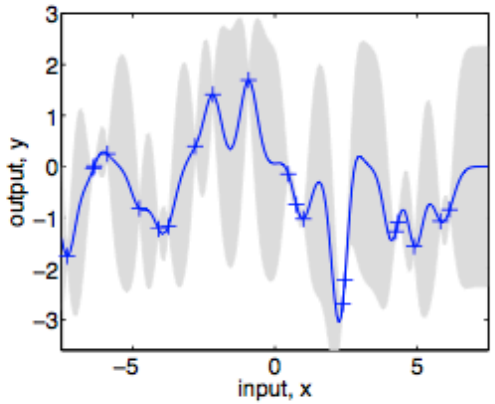
Efectos de los Hiperparámetros



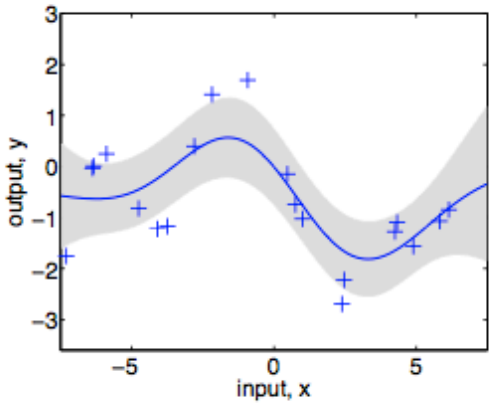
Efectos de los Hiperparámetros



(a), $\ell = 1$



(b), $\ell = 0.3$



(c), $\ell = 3$

Cálculo de los Hiperparámetros

- Recordando la definición de una distribución gaussiana multivariable:

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

- $P(y|x,\theta)$ sigue una distribución gaussiana multivariable con media cero y covarianza de $K+\sigma_n^2 I$ (si consideramos ruido)

$$logp(y|x,\theta) = -\frac{1}{2}y^T(K+\sigma_n^2 I)^{-1}y - \frac{1}{2}log|K+\sigma_n^2 I| - \frac{n}{2}log2\pi$$

Hiperparámetros

- Los hiperparámetros nos determinan las posibles distribuciones
- Para obtener los hiperparámetros podemos derivar con respecto a θ, pero antes es importante usar las siguientes dos expresiones:

$$\frac{\partial}{\partial\theta}K^{-1} = -K^{-1}\frac{\partial K}{\partial\theta}K^{-1}$$

donde $\frac{\partial K}{\partial\theta}$ es una matriz con las derivadas de sus elementos.

$$\frac{\partial}{\partial\theta}log|K| = tr(K^{-1}\frac{\partial K}{\partial\theta})$$

donde *tr* o *trace* es la suma de los elementos de la diagonal de la matriz

Hiperparámetros

- Entonces:

$$\frac{\partial}{\partial \theta_j} p(y|x,\theta) = \frac{1}{2} y^T K^{-1} \frac{\partial K}{\partial \theta_j} K^{-1} y - \frac{1}{2} tr(K^{-1} \frac{\partial K}{\partial \theta_i})$$

$$= \frac{1}{2} tr((\alpha\alpha^T - K^{-1}) \frac{\partial K}{\partial \theta_i})$$

  donde $\alpha = K^{-1} y$

- Para obtener entonces los hiperparámetros se sigue un proceso basado en gradiente (es un problema de optimización no-convexo), por ejemplo, basado en gradiente conjugado o quasi-Newton
- Se puede caer en mínimos locales

Procesos Gaussianos

Limitaciones:

- La complejidad de la inferencia es $O(n^3)$ por la inversión de la matriz.
- No son capaces de lidear con discontinuidades

Procesos Gaussianos

En términos prácticos, para programar GPs necesitamos:

- Calcular la matriz de covarianza.
- Invertirla
  Esto se realiza normalmente usando la factorización de Cholesky.
  Básicamente toma una matriz simétrica y la descompone en el producto de una matriz triangular inferior L (el factor de Cholesky) y su traspuesta:
  $LL^T = K$

Procesos Gaussianos

- Determinar los valores de los hiperparámetros (proceso de optimización).
  Nuestra estimación *a posteriori* de θ ocurre cuando $p(\theta|x, y)$ es máximo, lo que corresponde a maximizar $log(p(y|x, \theta)$ dado por:

$$logp(y|x, \theta) = -\frac{1}{2}y^T K^{-1} - \frac{1}{2}log|K| - \frac{n}{2}log2\pi$$

  Lo cual se puede resolver usando un algoritmo de optimización multivariable como gradiente conjugado, Nelder-Mead simplex, etc.