Apuntes Gaussian Processes

Patricio Whittingslow

March 2019

1 Prefacio

Es importante la palabra lineal. Aquí no hay funciones con cuadrados ni nada por el estilo. Existe una función global lineal que toma como argumento las variables elegidas.

Glosario Términos

$\mathbf{x}$ Input vector ($\mathbb{D}$). Que variables eligo para modelo.

$\mathbf{y}$ Vector objetivo (n). Observaciones/mediciones.

$\mathbf{w}$ Weight vector ($\mathbb{D}$). Aqui van los parámetros a obtener de la regresión lineal. La "solución" de la regresión lineal.

$\mathbb{D}$ Dimensión del problema. Cuantas variables elegí para el modelo.

n Cantidad de observaciones.

X Matriz de diseño ($\mathbb{D} \times n$).

$\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ Ruido. Sigue una distribución gausiana independiente, idénticamente distribuida con promedio 0 y varianza σ_n^2

$|\mathbf{v}| = \left(\sum_i \mathbf{v}_i^2\right)^{\frac{1}{2}}$ Longitud euclidiana del vector $\mathbf{v}$.

σ_n Asumimos que las mediciones están alejadas de la función f por causa de ruido ε.

Σ_p Matriz de covarianza o **matriz de varianza-covarianza**.

Glosario Subindices

i, j, k, p, q Refiere a un elemento de un tensor, sea vector,matriz, etc.

1.1 Introducción

Dado datos de entrenamiento $(X, \mathbf{y})$ se quiere efectuar predicciones para nuevas entradas $\mathbf{x}_*$ usando una función f, por lo tanto, es claro que el problema es *inductivo*. Para lograr esto tenemos que suponer ciertas características sobre nuestra función f. Si no acotáramos f entonces cualquier función que sea consistente con los datos de entrenamiento valdría. Este es el problema que intenta resolver el aprendizaje de maquina, el **Machine Learning.**

2 Regresion Lineal

Para que sea mas didáctico la aplicación de la regresión lineal se va plantear un problema a resolver: *Se tiene un cilindro de diámetro D y largo L que enfrenta un fluido de viscosidad μ, densidad ρ y velocidad U_∞. Obtener una expresión para la fuerza F_D que el fluido ejerce sobre el cilindro?*

Cabe destacar que el problema no es lineal y que probablemente la regresión lineal que obtengamos sea valida para unos pocos puntos. Igual intentaremos hallar la regresión lineal.

Primero tenemos que partir de datos, preferiblemente **muchos** datos. Vamos a obviar la variable L, ya que si dividimos la fuerza por el largo obtenemos la fuerza por unidad de largo y reducimos la dimensión del problema sin perder información.

Abajo esta la matriz de diseño para el problema. Se efectuaron n observaciones o mediciones, por ende también se tienen n resultados para la fuerza $y = F_{D_i}$.

$$X = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_n \\ \rho_1 & \rho_2 & \dots & \rho_n \\ U_1 & U_2 & \dots & U_n \\ D_1 & D_2 & \dots & D_n \end{bmatrix}$$

$$\mathbf{y} = \begin{Bmatrix} F_1 \\ F_2 \\ \vdots \\ F_n \end{Bmatrix}$$

2.1 No-Bayesiana

Como las mediciones nunca son exactas (siempre hay error humano o por el sistema de medición) se puede proponer ruido (ε) de distribución gaussiana $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ que se suma a nuestra regresión para obtener los valores de y tal que

$$f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{w} \qquad\qquad y = f(\mathbf{x}) + \varepsilon$$

Se presentaron dos nuevos vectores al usuario, $\mathbf{x}$ y $\mathbf{w}$, los vectores *input* y de *pesos* (*weight* en inglés), respectivamente. Ambos tienen dimensión del problema, en este caso $\mathbb{D} = 4$.

$$\mathbf{x} = \begin{Bmatrix} \mu_i \\ \rho_i \\ U_i \\ D_i \end{Bmatrix} \qquad \mathbf{w} = \begin{Bmatrix} w_a \\ w_b \\ w_c \\ w_d \end{Bmatrix}$$

Behold. La regresión lineal entonces quedaría de la forma

$$f(\mathbf{x}) = w_a\mu_i + w_b\rho_i + w_cU_i + w_dD_i = y_i = F_i$$

Resolver el problema (efectivamente: obtener el vector $\mathbf{w}$) requiere un uso pesado de la probabilidad, en particular el teorema de Bayes. También se tiene que elegir un *prior*, es decir, dejar expresadas nuestras creencias de los parametros (pesos) antes de mirar las observaciones. Esto toma la forma de la matriz de covarianza Σ_p. Si suponemos que los parametros se distribuyen según $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$ entonces:[1]

$$p(\mathbf{w}|X, \mathbf{y}) \sim \mathcal{N}(\bar{\mathbf{w}}, A)$$

$$A = \left(\sigma_n^{-2} X X^{\mathrm{T}} + \Sigma_p^{-1}\right)^{-1} \tag{1}$$

$$\bar{\mathbf{w}} = \sigma_n^{-2} \left(\sigma_n^{-2} X X^{\mathrm{T}} + \Sigma_p^{-1}\right)^{-1} X\mathbf{y} \tag{2}$$

Donde $\bar{\mathbf{w}}$ son los promedios de los pesos hallados. A es la *matriz de covarianza de la distribución posterior.*

Se suele referir al promedio de los pesos $\bar{\mathbf{w}}$ como MAP, *maximum a posteriori.*

2.2 Bayesiana

Se ha hablado de la formulación No-Bayesiana hasta ahora. De que se trata esto de si es Bayesiano o no? En el la formulación Bayesiana el caso prueba tiene la forma del vector de entrada, y la regresión se calcula para ese vector prueba. Por ende, en el esquema Bayesiano no existe un único vector $\mathbf{w}$. Un subíndice $_*$ se refiere a un caso prueba.

$$p(f_*|\mathbf{x}_*, X, \mathbf{y}) = \mathcal{N}\left(\sigma_n^{-2}\mathbf{x}_*^{\mathrm{T}} A X\mathbf{y}, \mathbf{x}_*^{\mathrm{T}} A \mathbf{x}_*\right) \tag{3}$$

Se usará esta formulación para la regresión en el espacio funcional. A se definió anteriormente.

3 Regresión en el espacio funcional

Si bien hemos obtenido una regresión cuando calculamos $\bar{\mathbf{w}}$, no resuelve el problema debido a que el problema *no es lineal.*

[1] Aún no se ha explicado como se calcula Σ_p.

Esto se resuelvo proyectando las variables de entrada a un espacio funcional. Nuestra función para la regresión tendría la forma

$$f(\mathbf{x}) = \phi(\mathbf{x})^{\mathrm{T}}\mathbf{w} \tag{4}$$

esta función ϕ mapea el vector de entrada $\mathbf{x}$ de un espacio $\mathbb{D}$ a un espacio funcional N. La aplicación de este modelo es análoga a la regresión lineal excepto que donde antes teníamos X ahora tenemos:

$$\Phi = \Phi(X)$$

$$p(f_*|\mathbf{x}_*, X, \mathbf{y}) \sim \mathcal{N}\left(\sigma_n^{-2}\phi(\mathbf{x}_*)^{\mathrm{T}}A\Phi\mathbf{y}, \phi(\mathbf{x}_*)^{\mathrm{T}}A\phi(\mathbf{x}_*)\right) \tag{5}$$

donde

$$A = \left(\sigma_n^{-2}\Phi\Phi^{\mathrm{T}} + \Sigma_p^{-1}\right)^{-1}$$

3.1 Kernel

Si uno quisiese ahorrar tiempo y no calcular la inversa de A (calculo computacional pesado) podria recurrir a la expresión (donde $\phi(x) = \phi$):

$$p(f_*|\mathbf{x}_*, X, \mathbf{y}) \sim \mathcal{N}\left(\phi_*^{\mathrm{T}}\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\mathbf{y}, \phi_*^{\mathrm{T}}\Sigma_p\phi_* - \phi_*^{\mathrm{T}}\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\Phi^{\mathrm{T}}\Sigma_p\phi_*\right) \tag{6}$$

donde $K = \Phi^T\Sigma_p\Phi$.

Note que en la ecuación 6 es prevalente el producto de el espacio funcional en las formas $\phi_*^{\mathrm{T}}\Sigma_p\Phi$, $\phi_*^{\mathrm{T}}\Sigma_p\phi_*$ y $\Phi^{\mathrm{T}}\Sigma_p\Phi$. Definiendo entonces una función $\psi(\mathbf{x}) = \Sigma_p^{\frac{1}{2}}\phi(\mathbf{x})$ obtenemos

$$k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x}) \cdot \psi(\mathbf{x}') \tag{7}$$

donde $\mathbf{x}$ y $\mathbf{x}'$ pueden ser las variables de entrada o casos prueba (denotado con subíndice $*$). Se le suele decir la función kernel[2] a k y es de sumo interés en la rama del aprendizaje de maquina. Si reescribieramos 6:

$$p \sim \mathcal{N}\left(k(\mathbf{x}_*, X)(k(X, X) + \sigma_n^2 I)^{-1}\mathbf{y}, k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathbf{x}_*, X)(k(X, X) + \sigma_n^2 I)^{-1}k(X, \mathbf{x}_*)\right) \tag{8}$$

Puede ser que llegado a este punto el lector se sienta ansioso por no tener las herramientas de como aplicar 8. Para aliviar dichas dudas se presenta una posible función de covarianza, la *squared exponential* (SE).[3]

$$\mathrm{cov}(f(\mathbf{x}_p), f(\mathbf{x}_q)) = k(\mathbf{x}_p, \mathbf{x}_q) = \exp(-\tfrac{1}{2}\left|\mathbf{x}_p - \mathbf{x}_q\right|^2/\ell) \tag{9}$$

donde ℓ es la longitud característica del proceso gausiano.

[2]También llamada función covarianza.

[3]también llamada la gausiana o *Radial Basis Function*

Definición 3.1 *Un* Proceso Gausiano *es una colección de cualquier número finito de variables aleatorias, todas distribuidas gausianamente.*

3.2 Prediccion con ruido

Suponiendo que $y = f(x) + \varepsilon$ siendo el ruido $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$

$$\operatorname{cov}(y_p, y_q) = k(\mathbf{x}_p, \mathbf{x}_q) + \sigma_n^2 \delta_{pq} \qquad o \qquad \operatorname{cov}(\mathbf{y}) = K(X, X) + \sigma_n^2 + I$$

La distribuci'on de $\mathbf{y}$ y las salidas según el *prior* elegido son

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