Regresión Kernel por mínimos cuadrados: interpretación estadística, procesos gaussianos

Antonio Sala Piqueras

Identificación de sistemas complejos

Dept. Ing. Sistemas y Automatica (DISA)

Universitat Politècnica de València (UPV)

Video-presentación disponible en:

http://personales.upv.es/asala/YT/V/**kstat**.html

Presentación

Motivación:
La regresión lineal clásica tiene una interpretación estadística (mínimos cuadrados $\Leftrightarrow$ mínima varianza). Si la versión Kernel es una reformulación equivalente, también debería tenerla.

Objetivos:
Comprender la interpretación estadística de la regresión regularizada Kernel. Comprender la idea de interpolación como realización de un proceso estocástico multidimensional.

Contenidos:
Revisión: kernel ridge regression y mejor predicción estadística. Similitudes entre las fórmulas. Procesos gaussianos, formalismo y ecuación de predicción. Conclusiones.

Revisión: Mínimos cuadrados regularizados

Regresión contraída (Ridge Regression), versión Kernel:
Consideramos predecir una variable y a partir de q características (regresores, $\phi_{q\times 1}(x)$) con modelo $\hat{y} = \theta_{1\times q} \cdot \phi_{q\times 1}(x)$, con $X_{q\times N} := [\phi(x_1) \ \dots \ \phi(x_N)]$, $Y_{1\times N} := [y_1 \ \dots \ y_N]$

$$J(\theta) = c\|\theta\|^2 + \sum_{i=1}^{N}(y_i - \theta\phi(x_i))^2 = c\|\theta\|^2 + (Y - \theta X)\cdot(y - \theta X)^T$$

Su sol. óptima coincide, cambiando $w_{1\times N} \cdot X^T = \theta$, $K_{N\times N} = X^T X$ (simétrica), con la de:

$$\begin{aligned} J(w) &= c(wKw^T) + (Y_{1\times N} - w_{1\times N}K_{N\times N})\cdot(Y - wK)^T \\ &= YY^T - YKw^T - wKY^T + w(K^2 + cK)w^T \end{aligned}$$

Revisión: Parámetro w óptimo

Del mismo modo que en el caso de los mínimos cuadrados "clásicos" (solución vía matriz pseudoinversa), el coste es cuadrático en w. Con lo que:

$$w_{1\times N}{}^{*} = Y_{1\times N} \cdot (K_{N\times N} + c \cdot I_{N\times N})^{-1}$$

En el caso de múltiples variables a predecir,

$$w_{m\times N}{}^{*} = Y_{m\times N} \cdot (K_{N\times N} + c \cdot I_{N\times N})^{-1}$$

Revisión: construcción del aproximador

Si no se dispone de las funciones $\phi(x) \in \mathbb{R}^q$ cuyos productos escalares generan el Kernel, pero se dispone de una función generadora $\kappa(x, z)$, $K_{ij} = \kappa(x_i, x_j)$, que abreviamos con $K = \kappa(X, X)$, el modelo de regresión ante un nuevo x genérico se reconstruye:

$$\hat{y}(x) = w^* \kappa(X, x)$$

donde $\kappa(X, x) \in \mathbb{R}^{N \times 1}$ tiene el elemento i dado por $\kappa(x_i, x)$.
Ello resulta en un modelo "interpolativo":

$$\hat{y}(x)_{m \times 1} = Y_{m \times N} \cdot (\kappa(X, X) + cI)_{N \times N}{}^{-1} \cdot \kappa(X, x)_{N \times 1}$$

Interpretación estadística: preliminares

La mejor predicción lineal (mínima vza.) de unos datos de media cero "$a_{m\times 1}$" a partir de otros "$b_{r\times 1}$", es

$$\hat{a} = \Sigma_{ab,m\times r}{\Sigma_{b,r\times r}}^{-1} \cdot b_{r\times 1}$$

y, si son vectores fila $\bar{a}_{1\times m}$ a partir de $\bar{b}_{1\times r}$:

$$\hat{\bar{a}}_{1\times m} = \bar{b}_{1\times r} \cdot {\Sigma_{b,r\times r}}^{-1}\Sigma_{ba,r\times m}$$

Si se añade un ruido de medida en la información b, $b_{noise} = b + v$, entonces, la matriz VC de (a, b_{noise}) es:

$$\begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b + V \end{pmatrix}$$

siendo V la vza de v. mejor predición de $a_{1\times m}$ dado b_{noise}:

$$\hat{a} = b_{noise} * (\Sigma_b + V)^{-1}\Sigma_{ba}$$

Hagamos la conexión conceptual

$\hat{a} = b_{noise} * (\Sigma_b + V)^{-1}\Sigma_{ba}$ $\qquad$ $\hat{y}(x)_{m\times 1} = Y_{m\times N} \cdot (\kappa(X,X) + cI)_{N\times N}{}^{-1} \cdot \kappa(X,x)_{N\times 1}$

Se puede hacer una analogía si:

- $K = \kappa(X,X)$ lo identificamos como "Σ_b", la matriz de varianzas-covarianzas entre los datos "limpios".
- $c \cdot I$ lo interpretamos como vza. ruido de medida V que contamina las muestras de la variable disponible para la predicción, Y.
- $\kappa(X,x)$ lo interpretamos como la matriz de covarianza entre la variable a predecir y las variables medidas.

¡Eureka!: se puede interpretar como predicción estadística.

Procesos estocásticos multidimensionales

Consideremos que queremos estimar un modelo $\hat{y} = f(x)$ con $x \in \mathbb{R}^n$.

Concibamos **función** f como un **conjunto infinito de variables aleatorias** $f(x) \sim N(0, \Sigma)$.

Suponemos una cierta "suavidad" de la función: en términos estadísticos, que la correlación estadística entre dos de esas variables depende de la distancia entre sus x asociadas.

***Analogía "temporal"** (ruido coloreado): correlación entre $x(t)$ y $x(t+\tau)$ mayor cuanto más pequeño sea τ. Función de autocorrelación $->$ Power spectral density, salida de una cierta FdT ante ruido blanco, etc.

- Generalizamos un proceso estocástico "temporal" (infinitas vbles. aleatorias asociadas cada una a un instante de tiempo) a más dimensiones (infinitas vbles. aleatorias asociadas, cada una, a un elemento x de $\mathbb{R}^n$).

Función de covarianza

Supongamos que tenemos un proceso (colec. ∞ vbles. aleat.) sobre $x \in \mathbb{R}^n$:

- **Modelo estadístico:** Supongamos que $cov(f(x), f(z)) = \kappa(x, z)$, siendo $\kappa : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ una función conocida.

  Obviamente, la varianza de x sería $\kappa(x, x)$ y la correlación entre $f(x)$ y $f(z)$ será $\rho_{f(x),f(z)} = \frac{\kappa(x,z)}{\sqrt{\kappa(x,x)\kappa(z,z)}}$

- **Observaciones:** Supongamos que disponemos de medidas con ruido del valor de la función en un conjunto de puntos $y_i = f(x_i) + w_i$, $i = 1, \ldots N$. El ruido de medida es no correlado entre diferentes x, tiene varianza c en cada punto ("ruido blanco" i.i.d.).

Ecuación de predicción

1. La matriz VC de la información disponible (y_i) es $\Sigma_{info} = \kappa(X, X) + c \cdot I$.
2. La matiz de covarianza entre los $f(x_i)$ y un cierto $f(x)$ es $\kappa(X, x)$.
3. La mejor predicción lineal (mínima vza.) de $f(x)$ dado Y es:
$$\hat{f}(x) = Y \cdot (\kappa(X, X) + c \cdot I)^{-1} \cdot \kappa(X, x)$$
4. La varianza del error de predicción es:
$$E((f(x) - \hat{f}(x)^2) = \kappa(x, x) - \kappa(X, x)^T (\kappa(X, X) + c \cdot I)^{-1} \kappa(X, x)$$

*Si el proceso es **gaussiano** (todas las vbles. distribución normal), esa fórmula sirve para determinar intervalos de confianza.

Relación con filtrado en sistemas dinámicos

En el caso $x \in \mathbb{R}$ el resultado es totalmente análogo al caso de proceso estocástico "temporal".

- Si la función $\kappa(x, y)$ depende de la distancia $\kappa(x, y) = \widetilde{\kappa}(\|x - y\|)$, sería el análogo a la función de "autocovarianza" de un proceso estocástico $V_\tau = E(x(t)x(t - \tau))$.
- La **transf. Fourier** de $\widetilde{\kappa}$ sería la **power spectral density** (PSD) del proceso, su factorización espectral ($\approx$ raíz cuadrada) sería la FdT que, tras pasar un ruido blanco por ella, generaría dicha PSD.

Ejemplo: Kernel $k(x, y) = e^{-\|x-y\|}$ se interpreta como $R_\tau = e^{-|\tau|}$, con PSD $\frac{1}{w^2+1}$, generada por el paso de un ruido blanco por el filtro $G(s) = \frac{1}{s+1}$.

Conclusiones

- Del mismo modo que la regresión lineal clásica tiene una interpretación estadística, también la tiene su versión Kernel.
- La interpretación es como un conjunto infinito de variables aleatorias $f(x)$, proceso estocástico, de las que se dispone de un conjunto de medidas con ruido en distintos puntos y de un modelo de la covarianza o correlación en función de la distancia... o cualquier $\kappa(x, y)$ que se suponga que describe la función en la aplicación.
- Permite, suponiendo una estructura subyacente de tipo "Gaussian Process", estimar intervalos de confianza de las variables predichas.