

Estadística noparamétrica. Trabajo en grupo

Máster en Técnicas Estadísticas. Curso 2009-2010

Selección del parámetro de suavizado en estimación no paramétrica de la densidad y de la regresión

El objetivo de este trabajo es comprobar el funcionamiento los métodos de selección del parámetro de suavizado vistos en clase, tanto para el estimador tipo núcleo de la densidad como para el estimador local lineal de la regresión. Para validar su comportamiento se realizará un estudio de simulación y un análisis de un conjunto de datos reales. El conjunto de datos a emplear será el data frame denominado `airquality`, que contiene diversas medidas de la calidad del aire en Nueva York entre Mayo y Septiembre de 1973. El análisis se centrará en las medidas de temperatura y niveles de ozono.

Estimación de la densidad

En el caso de la densidad, cada grupo implementará en R uno de los siguientes tres métodos:

1. Validación Cruzada
2. Validación Cruzada Sesgada
3. Método plug-in o de Seather y Jones

En esta parte se realizará un pequeño estudio de simulación. En todo el estudio se utilizará el núcleo gaussiano. Como *competidor* del método analizado se considerará la ventana normal, \hat{h}_{NS} , basada en estimar la ventana AMISE suponiendo que la distribución de los datos es la normal. Como criterio de error se empleará el error cuadrático integrado

$$ISE(h) = \int (\hat{f}_{n,K}(x) - f(x))^2 dx.$$

Este criterio de error permitirá decidir cuál de las dos ventanas seleccionadas comete menos error. Como modelos de prueba se considerarán las 15 densidades descritas en Marron y

Wand (1992). Estas densidades se han convertido en un estándar para validar cualquier método de estimación de la densidad y están programadas en R. Las funciones necesarias para evaluar cada una de las densidades o simular muestras a partir de las mismas están en la librería `nor1mix`. Por ejemplo, el comando `x<-rnorMix(n,MW.nm1)` genera una muestra de tamaño n de la primera de las densidades de Marron y Wand (que es, como no, la normal estándar). La función `dnorMix` se usa para evaluar las funciones de densidad de cada modelo.

Para realizar el estudio de simulación se generarán $B = 500$ muestras de tamaño $n = 100$ de cada una de las 15 densidades de Marron y Wand. Para cada muestra se estimará la ventana h usando los dos métodos analizados y evaluaremos el error cometido. Así, para cada densidad, se tendrán dos series de B números que representan el error cometido por cada uno de los métodos. Para comparar estas series de números se deberán usar aquellos resúmenes gráficos y/o numéricos que parezcan más adecuados: comparación de medias, desviación típica, diagramas de cajas, porcentaje de veces que un método gana a otro, etc. Estos resúmenes deberían permitir extraer conclusiones sobre los dos métodos que se comparan. El objetivo final es la comparación de los métodos considerados.

Regresión y análisis de datos

El objetivo de esta parte es analizar la relación que existe entre la temperatura y la concentración de ozono en la ciudad de Nueva York, para lo que se emplearán los datos contenidos en el data frame `airquality`. La variable independiente será la temperatura. Para hacer el análisis de regresión se considerará el método método local lineal, abordando los siguientes problemas:

1. Analizar la distribución univariante de las dos variables involucradas en el estudio. ¿Qué se puede decir sobre el clima de Nueva York (al menos en la década de los 70)?
2. Escribir una función en R que, dada una muestra $\{(x_i, y_i) : i = 1, \dots, n\}$ y una ventana h , devuelva el valor del estimador local lineal en el vector $x = (x_1, \dots, x_n)$. ¿Cómo adaptar el código para que la función nos permita evaluar el estimador en un vector arbitrario t , no necesariamente igual a x ?
3. Escribir una función en R que permita calcular la función de validación cruzada para el estimador local lineal. ¿Qué ventana selecciona el criterio de validación cruzada para los datos analizados?

4. Dibujar los datos analizados y el ajuste realizado por el método local lineal, usando como parámetro de suavizado la ventana de validación cruzada. ¿Resulta satisfactorio el ajuste realizado?
5. En regresión también existe una alternativa plug-in a la selección del parámetro de suavizado, implementada dentro de la librería `KernSmooth`. La función `dpill` permite calcular la ventana plug-in para la regresión. Utilizar dicha función para seleccionar un parámetro de suavizado alternativo al calculado por validación cruzada. ¿Cuál de las dos alternativas parece más razonable? ¿Por qué?
6. Usar las rutinas programadas en clase para el estimador de Nadaraya-Watson para analizar los datos. ¿Dónde se observan mayores diferencias con el método local lineal?
7. A la vista de los resultados anteriores, ¿se podría considerar un modelo sencillo para la descripción de los datos? Utilizar la regresión lineal simple para estimar dicho modelo. Comparar los resultados de ese ajuste lineal con los obtenidos con las técnicas no paramétricas

Forma de presentar los resultados

Una vez completados los análisis se debe entregar una memoria, preferiblemente escrita en \LaTeX . Esta memoria deberá constar de las siguientes partes:

1. Introducción y objetivos.
2. Descripción de los métodos considerados en el estudio. Se debe presentar de forma precisa los algoritmos usados en cada caso, detallando todos los cálculos realizados para implementar el código R utilizado.
3. Descripción del estudio de simulación y presentación de resultados.
4. Análisis de los datos.
5. Bibliografía
6. Apéndice con el código utilizado.

Referencias

Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error. *The Annals of Statistics*, vol. 20, pp. 712–736.