Estimación kNN de la función de regresión

Beatriz Coronado Sanz
Andrea Prieto García

Máster Universitario en Matemáticas
Universidad de Sevilla

Modelado y Predicción Estadística

ÍNDICE

Introducción

Consideramos una variable aleatoria (v.a.) continua X con función de densidad (f.d.d.) f.

El objetivo es estimar f a partir de $X_1, ..., X_n$ de X. Este es el paso previo a estimar nuestro segundo y principal objetivo: estimar de forma no paramétrica $m(x)$ para tener $y = m(x) + \epsilon$.

En este trabajo nos centraremos en el denominado estimador de los k vecinos más cercanos (estimador k-NN).

Estimación kNN

El método k-NN fue descrito como una aproximación no paramétrica del análisis del discriminante por Cover y Hart (1967) o Fix y Hodges (1989).

Esta aproximación clasifica una característica desconocida de un objeto basándose en la similaridad total de los objetos conocidos que lo rodean.

Estimación kNN

El método de estimación de los k vecinos más cercanos es un caso particular de la estimación núcleo.

Una ventaja de este estimador es que la suavización varía acorde al número de observaciones en una región particular.

Estimador kNN

Sean $X_1, X_2, ..., X_n$ v.a. i.i.d. con f.d.d. $f(x)$. La densidad estimada del método de los k-vecinos más cercanos es (Orava, 2011)

$$\hat{f}_{kNN}(x,k) = \frac{1}{nr_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{r_n}\right) \tag{1}$$

donde $r_n = r_n(x)$ es la distancia euclídea entre x y el k-ésimo vecino más cercano a x entre los X_j.

Estimador kNN

K es la función núcleo que satisface $\int K(x)dx = 1$ y $k = \{k(n)\}$ el entero positivo que cumple $k \to \infty$ y $\frac{k}{n} \to 0$ cuando $n \to \infty$.

En este caso, utilizaremos el núcleo de Epanechnikov:

$$K(x) = \begin{cases} \frac{3}{4} & \text{si } x \in [-1,1] \\ 0 & \text{c.c.} \end{cases} = \frac{3}{4} I(|x| \leq 1)$$

El estimador $\hat{f}_{kNN}(x)$ mide la proximidad a los k elementos más cercanos a x. Si los k X_j están cerca, entonces $\hat{f}_{kNN}(x)$ será grande y si están lejos, entonces $\hat{f}_{kNN}(x)$ será pequeño.

Estimación kNN de la función de densidad

La función núcleo K es no negativa, simétrica respecto al origen y acotada.

La función de densidad que teníamos del estimador núcleo es una densidad en sí misma. Sin embargo, esto no es válido para el estimador k-NN: la integral de la función de densidad estimada del núcleo del estimador k-NN será cercana a 1, pero no es 1.

De nuevo, como ocurría con el estimador núcleo, la elección de la función núcleo K no tiene una gran influencia en el resultado final.

Estimación kNN de la función de densidad

Tenemos:

$$\hat{f}_{KNN}(x,k) = \frac{\text{nº de obs en } (x-r_n, x+r_n]}{\frac{4}{3}nr_n} = \frac{1}{nr_n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{r_n}\right)$$

con $K(x) = \frac{3}{4}I(|x| \leq 1)$.

El núcleo de Epanechnikov asigna peso $3/4$ a toda observación en la muestra que se encuentre a una distancia menor o igual a $r(n)$ de x y 0 al resto. No importa cuán cercanas o lejanas estén las $k-1$ observaciones restantes dentro de ese recinto (la observación k-ésima más cercana a x marca el valor de $r(n)$).

Estimación kNN de la función de densidad

Vamos a asignar un peso distinto a cada observación, dando más importancia a las k observaciones más cercanas:

$$\hat{f}_{KNN}(x,k) = \frac{1}{nr_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{r_n}\right) = \frac{1}{n}\sum_{i=1}^{n} K_{r_n}(x - X_i)$$

donde $K_{r_n}(\cdot) = \frac{1}{r_n}K(\frac{\cdot}{r_n})$ y K cumple $\int K(x)dx = 1$.

- K es la función núcleo.
- r_n es la ventana definida dependiendo de la k-ésima observación más cercana considerada.
- $\hat{f}_{KNN}$ es el estimador kNN de la f.d.d. $f(x)$.

Estimación kNN de la curva de regresión

El error cuadrático medio (MSE) de $\hat{m}(x)$ como estimador de $m(x)$ cumple $E[\hat{m}(x) - m(x)]^2 \to 0$ cuando $k \to \infty$ y $k/n \to 0$.

La construcción del estimador de los k vecinos más cercanos difiere de la construcción del estimador núcleo.

Mientras que el estimador núcleo $\hat{m}_h(x)$ fue definido como un promedio ponderado de las variables en un entorno fijado respecto a x (determinado en forma por el núcleo K y el tamaño de ventana h), el estimador de los k vecinos más cercanos es un promedio ponderado en un entorno variable de x.

Este entorno se define mediante las k variables X_j más cercanas a x en términos de distancia euclídea.

Estimación kNN de la curva de regresión

Definimos (Härdle, 1994), en este caso, el estimador de Nadaraya-Watson como

$$\hat{m}_k(x) = \frac{1}{n}\sum_{i=1}^{n} W_{k,i}(x)\, Y_i$$

donde $\{W_{k,i}(x)\}_{i=1}^{n}$ es la secuencia de pesos definida mediante el conjunto de índices $J_x = \{i : X_i \text{ es una de las } k \text{ obs más cercanas a } x\}$.

Con este conjunto de índices de las observaciones más cercanas, se definen los pesos asignados a cada x mediante:

$$W_{k,i}(x) = \begin{cases} n/k & \text{si } i \in J_x \\ 0 & \text{cc} \end{cases} \qquad \text{(Pesos Uniformes)}$$

El estimador de Nadaraya-Watson se corresponde con el ajuste constante local de mínimos cuadrados.

Estimación kNN de la curva de regresión

En un experimento en el que las observaciones sean cogidas en una red equiespaciada, los pesos del método k-NN son equivalentes a los pesos del estimador núcleo.

Sea $k = 2nh$, vamos a comparar $\{W_{k,i}(x)\}$ con $\{W_{h,i}(x)\}$ para un núcleo uniforme $K_u(x) = \frac{1}{2} I(|x| \leq 1)$ para x no cercana a los extremos.

En el fondo, para $i \in J_x$ se tiene que

$$W_{k,i}(x) = \frac{n}{2nh} = \frac{1}{2} h^{-1} = W_{h,i}(x)$$

En otras palabras, el estimador núcleo y el estimador k-NN coinciden para este caso.

Estimación kNN de la curva de regresión

Nótese que $\hat{m}(x)$ es una media ponderada de los valores de la variable respuesta, Y. Estas ponderaciones dependen de la variable independiente o predictora, X, y del parámetro k.

Sea n fijo, entonces:

- Si $k = n$, entonces $W_{k,i} = 1$ y, por tanto, $\hat{m}(x) = \overline{Y}$
- Si $k = 1$, entonces $W_{k,i} = n$, luego $x = X_i$ y, por tanto, $\hat{m}(X_i) = Y_i$

El el caso límite $k = 1$, en el cual las observaciones son reproducidas como X_i, si tenemos una x entre dos variables predictoras a la misma distancia, $\hat{m}(x)$ sería la media de los dos X_i.

Estimación kNN de la curva de regresión

Como en el caso del estimador núcleo, volvemos a tener un parámetro cuya función es suavizar la curva a predecir: k tiene que elegirse como función de los datos o de su tamaño, n .

El primer objetivo consiste en reducir la varianza considerando $k = k_n$ $(n \to \infty)$ como función del tamaño de muestra. Y el segundo consiste en mantener el error de aproximación lo más pequeño posible.

Este segundo fin se consigue si el conjunto de puntos cercanos a x se reduce asintóticamente a 0. Esto puede hacerse definiendo $k = k_n$ tal que $k_n/n \to 0$. Desafortunadamente, esta condición entra en conflicto con el primer objetivo.

Por tanto, nuestro fin será mantener la varianza tan pequeña como sea posible y, simultáneamente, elegir el máximo valor de k.

Estimación kNN de la curva de regresión

Resumiendo: tenemos un problema en donde hay que compensar entre una "buena aproximación" para la función de regresión y una "buena reducción" del ruido observado.

Este problema puede ser expresado formalmente como una expansión del MSE para el estimador k-NN.

Propiedades kNN

Teorema 1 (Lai, 1977) Sea $k \to \infty$, $k/n \to 0$, $n \to \infty$. Entonces, mediante los pesos definidos anteriormente se tiene que

$$E[\hat{m}(x)] - m(x) \approx \frac{1}{24 f(x)^3}[(m''(x)f(x) + 2m'(x)f'(x)](k/n)^2$$

$$var(\hat{m}(x)) \approx \frac{\sigma^2(x)}{k}$$

Luego, el valor óptimo para compensar la varianza y la esperanza sería tomar $k \sim n^{4/5}$.

Propiedades kNN

Stone (1977) propuso esta otra secuencia de pesos llamados *triangular and quadratic k-NN weights*

$$W_R(x) = \frac{K_R(x - X_i)}{\hat{f}_R(x)}$$

donde

$$\hat{f}_R(x) = \frac{1}{n} \sum_{i=1}^{n} K_R(x - X_i)$$

es la función de densidad estimada de $f(x)$ y,

$$K_R(u) = R^{-1} K(u/R)$$

donde $R = r_n$ definido como antes.

Propiedades kNN

Teorema 2 (Mack & Rosenblatt, 1981) Sea la función de densidad f acotada y sea la función núcleo tal que

$$\int |x|^2 K(x)dx < \infty$$

con $x \in \mathbb{R}^p$ Además, sea x con $f(x) > 0$ y f dos veces diferenciable en un entorno de x. Entonces, si

$$k = k(n) \to \infty \ , \ k/n \to 0 \text{ si } n \to \infty$$

se tiene

$$var(\hat{m}(x)) = \frac{f^2(x)}{k} \frac{\pi^{p/2}}{\Gamma\left(\frac{p+2}{2}\right)} \int K^2(u)du + o\left(\frac{1}{k}\right)$$

Con $\beta_2 = \int K^2(u)du$

Propiedades kNN. Teorema 3

Teorema 3 (Mack & Rosenblatt, 1981) Sea la función de densidad f acotada y sea la función núcleo tal que

$$\int |x|^2 |K(x)| dx < \infty, \quad \int |x| K(x) dx$$

con $x \in \mathbb{R}^p$. Además, sea x con $f(x) > 0$ y f dos veces diferenciable en un entorno de x.

Propiedades kNN. Teorema 3

Entonces, si

$$k = k(n) \to \infty \ , \ k/n \to 0 \text{ si } n \to \infty$$

se tiene

$$E[\hat{m}(x)] = m(x) + \frac{\left(\Gamma\left(\frac{p+2}{2}\right)\right)^{2/p}}{2\pi f(x)^{2/p}} \mu_2(K)\mathcal{H}(x)\left(\frac{k}{n}\right)^{2/p}$$
$$+ o_p\left(\left(\frac{k}{n}\right)^{2/p} + \frac{1}{k}\right)$$
$$+ \frac{e_1' N^{-1}}{2} \int_{D_x} [1\, u]' K(u) u' H^{1/2} \mathcal{H}(x) H^{1/2} u du + o_p\left(\text{tr}(H)\right)$$

Demostración del Teorema 3. Notación previa

Definimos previamente las siguientes matrices:

$$\mathrm{M} = E[Y|X] = \begin{pmatrix} E[Y_1|X_1] \\ \cdots \\ E[Y_n|X_n] \end{pmatrix} = \begin{pmatrix} m(X_1) \\ \cdots \\ m(X_n) \end{pmatrix}$$

con $E[Y_1|X_1, ..., X_n] = E[Y_1|X_1]$ e $(Y_1, X_1), ..., (Y_n, X_n)$ independientes.

$$e_1 = (1\mathbf{0}) \in \mathbb{R}^n$$

$$X = \begin{pmatrix} 1 & (X_1 - x)' \\ 1 & (X_2 - x)' \\ \cdots & \\ 1 & (X_n - x)' \end{pmatrix}$$

Demostración del Teorema 3. Notación previa

Definimos:

$$W = \text{diag}\left\{K(X_1 - x), ..., K(X_n - x)\right\}$$

siendo K la última función núcleo definida.

$$\mathcal{H}(x) \text{ matriz hessiana de } m$$

$$H = \text{diag}\left\{R^2\right\}$$

Demostración del Teorema 3. Notación previa

Definimos:

$$\mu_2(f) = \int x^2 f(x) dx \tag{2}$$

$$D_m(x) = \left(\frac{\partial m}{\partial x_1}, ..., \frac{\partial m}{\partial x_p} \right)'$$

$$Q_m(x) = [(X_1 - x)' \mathcal{H}(x)(X_1 - x), ..., (X_n - x)' \mathcal{H}(x)(X_n - x)]' \quad ((2))$$

$$D_x = \left\{ z : (x + H^{1/2} z) \in \text{sop}(f) \right\} \cap \text{sop}(K) \tag{D}$$

Demostración del Teorema 3. Notación previa

Definimos:

$$N = \int_{D_x} [1\ u]'[1\ u]K(u)du = \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix} \tag{N}$$

$$A = \begin{pmatrix} 1 & 0 \\ 0 & H^{1/2} \end{pmatrix} \tag{3}$$

Probaremos primero la igualdad para los puntos interiores y luego para los puntos de la frontera.

Demostración del Teorema 3

Como m es dos veces continuamente derivable, el teorema de Taylor implica que

$$M = X \cdot \begin{pmatrix} m(x) \\ D_m(x) \end{pmatrix} + \frac{1}{2}\mathcal{H}(x) \begin{pmatrix} (X_1 - x)^2 \\ \cdots \\ (X_n - x)^2 \end{pmatrix} + \tilde{R}(x)$$

Demostración Teorema 3

$$E[\hat{m}_{kNN}(x)|_{X_1=x,\ldots,X_n=x}] = e_1'(X'WX)^{-1}X'WM \qquad ((1))$$

Si desarrollamos ((1)) y tenemos en cuenta ((2)), obtenemos

$$E[\hat{m}_{kNN}(x) - m(x)|_{X_1=x,\ldots,X_n=x}] = \frac{1}{2}e_1'(X'WX)^{-1}X'W\{Q(x) + r\}$$

expresión en la que no aparece $D_m(x)$ ya que

$$e_1'(X'WX)^{-1}X'WX\begin{pmatrix} m(x) \\ D_m(x) \end{pmatrix} = m(x)$$

Demostración Teorema 3

Ahora, $n^{-1}X'WX =$

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} K(X_i - x) & \frac{1}{n}\sum_{i=1}^{n} K(X_i - x)(X_i - x)' \\ \frac{1}{n}\sum_{i=1}^{n} K(X_i - x)(X_i - x) & \frac{1}{n}\sum_{i=1}^{n} K(X_i - x)(X_i - x)(X_i - x)' \end{pmatrix} \quad ((3))$$

Demostración Teorema 3

Utilizando los resultados de la estimación de la densidad:

$$\frac{1}{n}\sum_{i=1}^{n} K(X_i - x) = f(x) + o_p(1)$$

$$\frac{1}{n}\sum_{i=1}^{n} K(X_i - x)(X_i - x) = \mu_2(K)HD_f(x) + o_p(H\mathbf{1})$$

$$\frac{1}{n}\sum_{i=1}^{n} K(X_i - x)(X_i - x)(X_i - x)' = \mu_2(K)f(x)H + o_p(H)$$

Demostración Teorema 3

Luego, $(n^{-1}X'WX)^{-1} =$

$$= \begin{pmatrix} f(x)^{-1} + o_p(1) & -D_f(x)'f(x)^{-2} + o_p(\mathbf{1}) \\ -D_f(x)f(x)^{-2} + o_p(\mathbf{1}) & \{\mu_2(K)f(x)H\}^{-1} + o_p(H^{-1}) \end{pmatrix}$$

Demostración Teorema 3

De la misma forma, vemos que $n^{-1}X'WQ(x) =$

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} K(X_i - x)(X_i - x)'\mathcal{H}(x)(X_i - x) \\ \\ \frac{1}{n}\sum_{i=1}^{n} \{K(X_i - x)(X_i - x)'\mathcal{H}(x)(X_i - x)\}(X_i - x) \end{pmatrix}$$

y $\tilde{R}$ al ser multiplicado por $e_1'(X'WX)^{-1}X'W$ pasa a ser de orden $o_p(\mathrm{tr}(H))$

Demostración Teorema 3

$$\frac{1}{n}\sum_{i=1}^{n}\left\{K(X_i-x)(X_i-x)'\mathcal{H}(x)(X_i-x)\right\}(X_i-x)=$$

$$=\int K(u)\left\{\left(H^{1/2}u\right)'\mathcal{H}(x)\left(H^{1/2}u\right)\right\}\left(H^{1/2}u\right)f(x+H^{1/2}u)du+$$

$$+o_p(H^{3/2}\mathbf{1})$$

$$=o_p\left(H^{3/2}\mathbf{1}\right)$$

Demostración Teorema 3

Se sigue por tanto que, $E[\hat{m}(x)|X_1, ..., X_n] - m(x) =$

$$= \frac{1}{2} f(x)^{-1} E\left[\frac{1}{n}\sum_{i=1}^{n} K(X_i - x)(X_i - x)'\mathcal{H}(x)(X_i - x)\right]$$
$$+ o_p\left\{\text{tr}(H)\right\}$$
$$= \frac{1}{2}\text{tr}\left\{H^{1/2}\mathcal{H}(x)H^{1/2}\int K(u)uu'du\right\} + o_p\left\{\text{tr}(H)\right\}$$
$$= \frac{1}{2}\mu_2(K)\text{tr}\left\{H\mathcal{H}(x)\right\} + o_p\left\{\text{tr}(H)\right\}$$

Demostración Teorema 3

Finalmente, $E[\hat{m}_{kNN}(x)|_{X_1=x,\ldots,X_n=x}] - m(x) =$

$$= \frac{1}{2}\mu_2(K)\mathrm{tr}\left(H\mathcal{H}(x)\right) + o_p\left\{\mathrm{tr}(H)\right\} =$$

$$= \frac{\left(\Gamma\left(\frac{p+2}{2}\right)\right)^{2/p}}{2\pi f(x)^{2/p}}\mu_2(K)\mathcal{H}(x)\left(\frac{k}{n}\right)^{2/p} + o_p\left(\left(\frac{k}{n}\right)^{2/p} + \frac{1}{n}\right)$$

Probaremos ahora la igualdad para los puntos de la frontera. Mediante un razonamiento análogo al anterior tomando

$$n^{-1}X'WX = f(x)ANA + o_p(A\mathbf{1}A)$$

$$n^{-1}X'WQ =$$

$$\begin{pmatrix} f(x)\mathrm{tr}\left\{H^{1/2}\mathcal{H}(x)H^{1/2}n_{22}\right\} + o_p\left\{\mathrm{tr}(H)\right\} \\ f(x)H^{1/2}\int_{D_x} uK(u)\left\{u'H^{1/2}\mathcal{H}(x)H^{1/2}u\right\}du + o_p\left\{H^{1/2}\mathbf{1}\mathrm{tr}(H)\right\} \end{pmatrix}$$

Por las hipótesis del Teorema 3, N es no singular luego

$$N^{-1} = \begin{pmatrix} n^{11}n^{12} \\ n^{21}n^{22} \end{pmatrix}$$

donde $n^{11} = (n_{11} - n_{12}n_{22}^{-1}n_{21})^{-1}$, $n^{12} = -(n_{12}/n_{11})n_{22}$ y $n_{22} = (n_{22} - n_{21}n_{12}/n_{11})^{-1}$

Resultando al final que

$$E[\hat{m}(x) - m(x)] =$$
$$= \frac{e_1' N^{-1}}{2} \int_{D_x} [1\, u]' K(u) u' H^{1/2} \mathcal{H}(x) H^{1/2} u du$$
$$+ o_p \{\mathrm{tr}(H)\}$$

Demostración Teorema 3

Para estimar el valor de la distribución de probabilidad, tomemos una región T suficientemente pequeña en la que se encuentre x, la probabilidad P de que un dato se encuentre en t está dada por

$$P = \int_T f(x)dx$$

en consecuencia, la probabilidad de encontrarnos k observaciones en T estará dada por una distribución binomial

$$B(k|n, f) = \binom{n}{k} P^k (1 - P)^{n-k}$$

es bien conocido que la esperanza de la distribución binomial es $E[k] = P \cdot n$. Por tanto, la esperanza de la fracción de puntos que se encuentran en la región T, k/n, está dada por $E[k/n] = P$.

Demostración Teorema 3

Vemos pues que haciendo n grande podemos centrar la distribución de fracción k/n en la media tanto como queramos y por tanto, tiene sentido la aproximación

$$k \approx n \cdot P$$

Si suponemos que la región T es pequeña y tiene un volumen V, podemos suponer que la distribución $f(x)$ es constante en T y aproximar

$$f(x) \approx \frac{k}{nV}$$

Demostración Teorema 3

En nuestro problema, el volumen será el formado por la hiperesfera de dimensión $p-1$ en p siendo el radio la distancia entre x y su vecino k-ésimo más cercano X_j. Por tanto:

$$V = \frac{\pi^{p/2} R^p}{\Gamma(p/2+1)}$$

Juntando ambas expresiones tenemos los elementos de $H^{-1/2}$:

$$R = \frac{\Gamma\left(\frac{p+2}{2}\right)^{1/p}}{\sqrt{\pi} f(x)^{1/p}} \left(\frac{k}{n}\right)^{1/p}$$

Propiedades kNN

En particular, para $x \in \mathbb{R}$ tenemos:

Teorema 4 (Mack y Rosenblatt) Sea $k \to \infty$, $\beta_2(f)$ y $\mu_2(f)$ definidos como en los *Teoremas 2 y 3*. Entonces:

$$E\left\{\hat{m}(x;k)|_{X=x}\right\} - m(x) \approx \left(\frac{k}{n}\right)^2 \frac{(m''f + 2m'f')(x)}{8f^3(x)}\mu_2(K) \qquad (4)$$

$$var\left\{\hat{m}(x;k)|_{X=x}\right\} = 2\frac{\sigma^2(x)}{k}\beta_2(K) \qquad (5)$$

Siendo $\sigma^2(x)$ la varianza de los datos.

Propiedades kNN

Como consecuencia del Teorema 3, al igual que ocurría para los pesos definidos en *Pesos Uniformes*, para conseguir un equilibrio entre la varianza y la esperanza k debe ser proporcional a $n^{4/5}$.

Comparamos el valor de la varianza y la esperanza en término de mínimos cuadrados del estimador de la curva de regresión mediante el método k-NN y el estimador núcleo para $\mathbb{R}$ en la siguiente tabla:

	Núcleo	k-NN
Esperanza	$h^2 \frac{(m''f + 2m'f)(x)}{2f(x)} \mu_2(x)$	$\left(\frac{k}{n}\right)^2 \frac{(m''f + 2m'f)(x)}{8f^3(x)} \mu_2(x)$
Varianza	$\frac{\sigma^2(x)}{nhf(x)} \beta_2(K)$	$\frac{2\sigma^2(x)}{k} \beta_2(K)$

Podemos observar que la igualdad se da en ambos métodos para $h = R = \frac{k}{2nf(x)}$ siendo R la distancia de x a su k-ésimo vecino más cercano.

Esto es porque un problema con el estimador k-NN puede verse como uno de núcleo siendo h la distancia entre x y su k-ésimo vecino más cercano. Además, en $\mathbb{R}$, el volumen de la hiperesfera de dimensión 0 es un intervalo:

$$V = \frac{\pi^{1/2}R}{\Gamma(\frac{1}{2}+1)} = \frac{\sqrt{\pi}R}{\sqrt{\pi}/2} = 2R$$

Y como sabemos que $f(x) = \frac{k}{nV} = \frac{k}{2nR}$, tenemos $R = \frac{k}{2nf(x)} = h$.

Elecciones

En el estimador núcleo vimos que la elección del núcleo K no es tan importante, pues cualquier núcleo razonable proporcionaba buenos resultados. Hemos elegido el núcleo de Epanechnikov porque con él se alcanza el mínimo AMISE.

A continuación discutiremos sobre la elección del parámetro k, pues determinará la ventana.

Criterios para evaluar la bondad de $\hat{f}$

Como el AMISE es

$$\text{AMISE}(\hat{f}) = \int \hat{var}(\hat{f}(x)) + \int \hat{E}^2(\hat{f})$$

usando las notaciones de la esperanza y la varianza tenemos:

$$\text{AMISE}(\hat{f_{kNN}}) = \frac{2}{k}\beta_2(K)b_2(f) + \frac{1}{64}\left(\frac{k}{n}\right)^4 \mu_2^2(K)\beta_2\left(\frac{f''}{f^2}\right) \qquad (6)$$

El valor de k que minimiza el AMISE($\hat{f}_{kNN}$) es

$$k_{\text{AMISE}} \approx \left(2n^{4/5}\left(\frac{\beta_2(K)\beta_2(f)}{\mu_2^2(K)\beta_2\left(\frac{f''}{f^2}\right)}\right)^{\frac{1}{5}}\right)$$

Criterios para evaluar la bondad de $\hat{f}$

El valor del k_{AMISE} será llamado k AMISE óptimo.

Como la expresión del k_{AMISE} depende de una función de densidad desconocida, no puede ser usado en la práctica.

Nuestro fin es sustituir la densidad desconocida f con una densidad de referencia que permita estimar correctamente el valor de k_{AMISE}.

El valor de $\beta_2(f) \to 0$ y entonces $k_{\text{AMISE}} \to \infty$.

Selección del número de vecinos en la práctica

Para la selección óptima del número de vecinos usaremos el criterio de validación cruzada:

$$k_{CV} = argmin_k \sum_{i=1}^{n} \{Y_i - \hat{m}_i(X_i)\}^2$$

Este método funcionará porque ya hemos visto que el estimador k-NN es igual al estimador núcleo siendo h la distancia entre un punto x y su k-ésimo vecino más cercano.

Modificar el número de vecinos para cada punto modificará la distancia h y nos permitirá elegir en la práctica un h óptimo para cada problema.

Análisis con R

El paquete **caret** incluye una serie de funciones que facilitan el uso de varios métodos complejos de clasificación y regresión. El comando más importante de este paquete es **train**, que permite aplicar un gran número de metodos de clasificación y regresión determinando los valores óptimos de sus parámetros mediante validación cruzada u otros métodos de remuestreo.

Para usar *train* es necesario elegir el método de clasificación que queremos usar (en nuestro caso **knn**), determinar los parámetros que queremos averiguar y fijar los datos sobre los que vamos a predecirlos y definir el método de remuestreo que se va a utilizar para determinar estos parámetros.

Análisis con R

Para el método de remuestreo usaremos un comando adicional proporcionado por el paquete *caret*: *trainControl*. Este comando nos indica con que método vamos a querer realizar el remuestreo (en nuestro caso **cv**). Por último, tenemos que indicar en el comando *train* el número por el que queremos dividir la muestra.

En resumen vamos a tener:

$$train(parametro, data = entrenamiento, method = \text{“}knn\text{”},$$

$$tuneLength = n^{\circ}div, trControl = trainControl(method = \text{“}cv\text{”}))$$

Análisis con R

Otro comando que usaremos, en este caso del paquete **stats**, será **predict**. Dado un objeto lineal, predice para unos nuevos datos su resultado en ese modelo. Usaremos este modelo para predecir el parámetro que queremos calcular en el conjunto de test (previamente calculado el parámetro en el conjunto de entrenamiento con **train**).

$$predict(modeloLineal, newData = test)$$

Ejemplo: vinos.csv

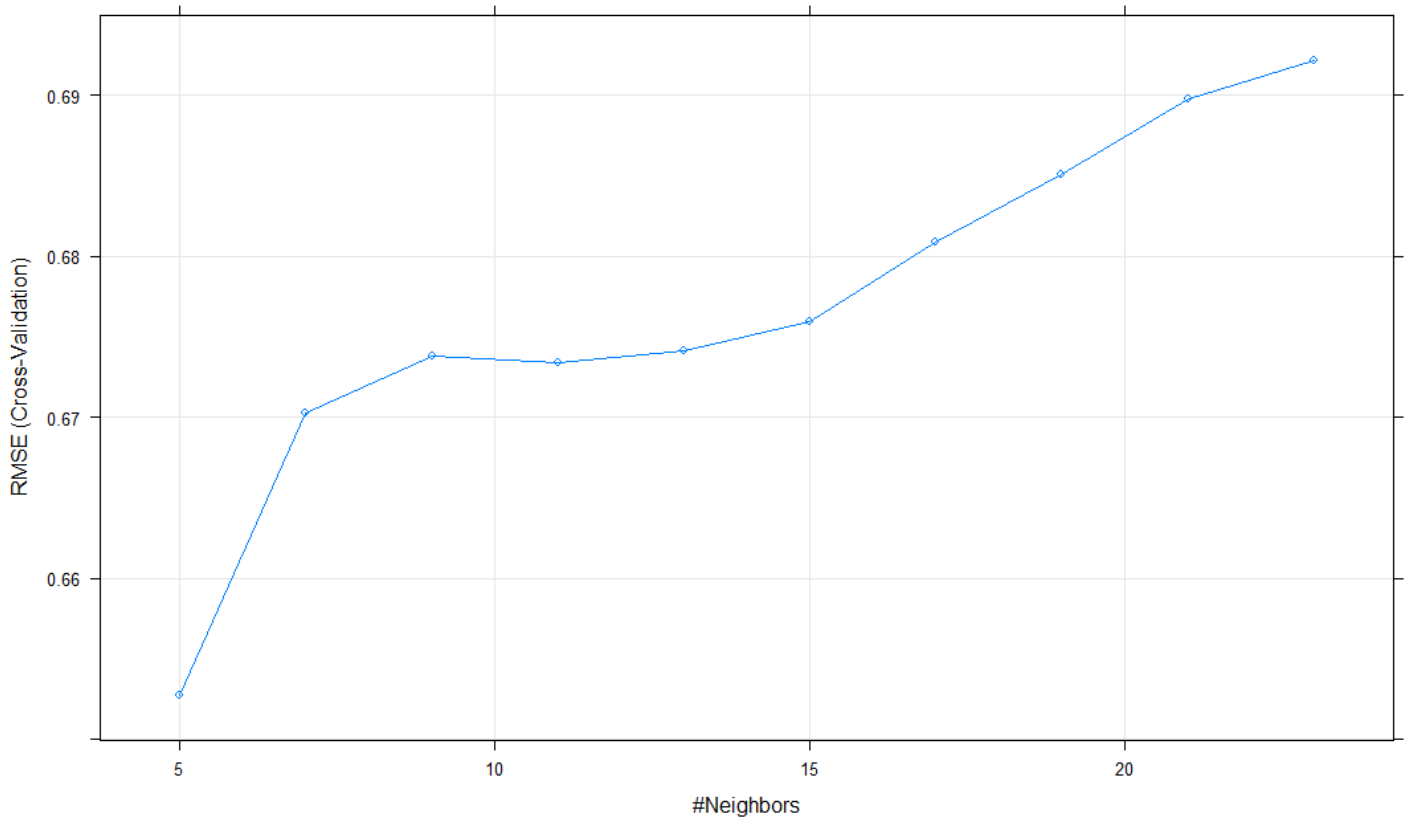
Vamos a usar un dataset de la página web Kaggle en donde se recogen 12 características de las variantes rojas del vino portugués "vinho verde". Algunas de estas características son el pH del vino, los sulfatos que tiene o la cantidad de alcohol que presenta.

Nosotras vamos a intentar predecir el alcohol en un conjunto de entrenamiento por el método de los k vecinos más cercanos. Luego usaremos el modelo calculado para predecir el alcohon en un conjunto de test y veremos el error que se ha cometido en la predicción.

El script con todo el código puede verse en el archivo *KnnRegression.R*.

Ejemplo: vinos.csv

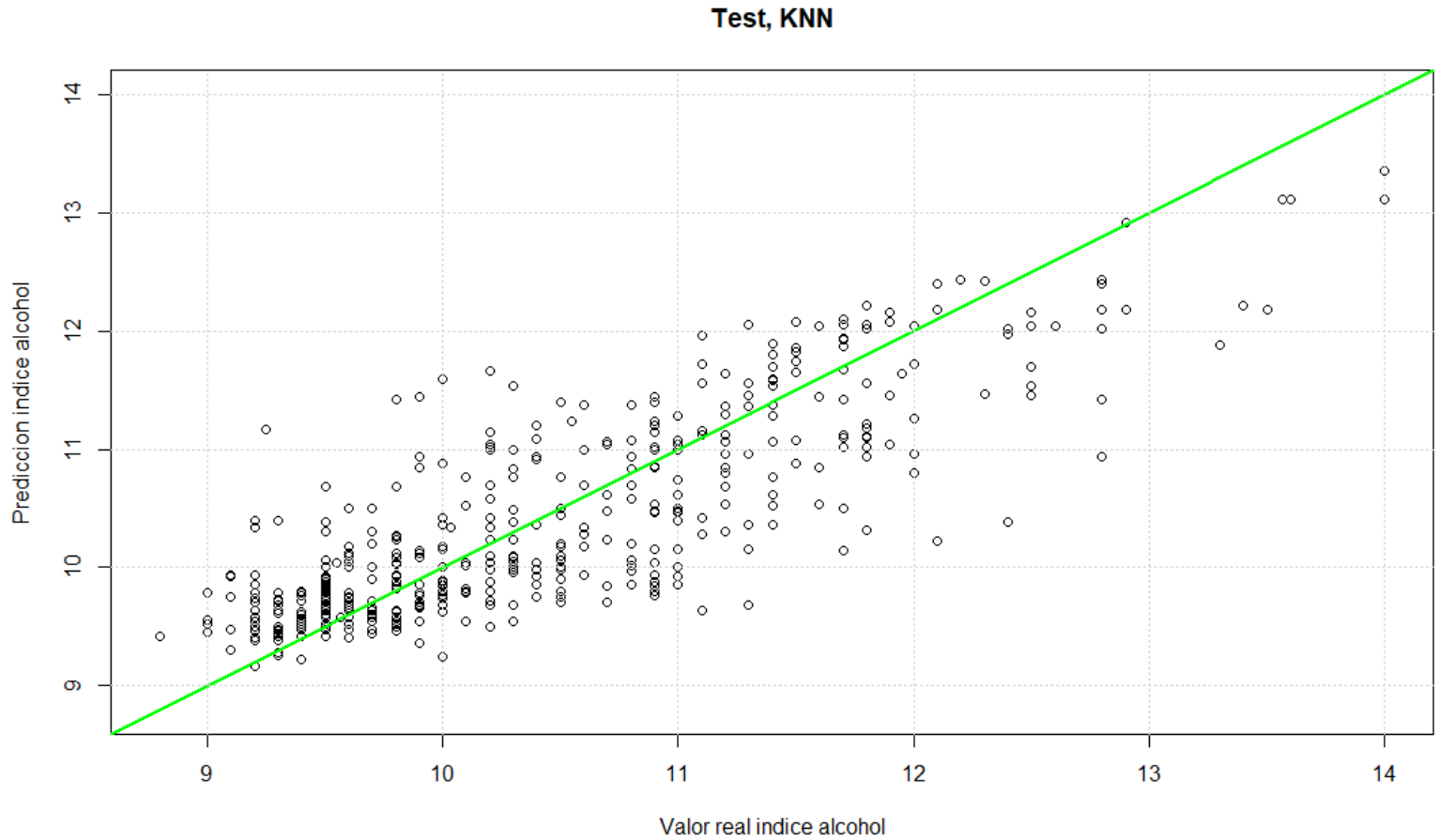
Al aplicar el comando **train** al conjunto de entrenamiento obtenemos la siguiente gráfica:



Observamos que el mínimo error para el RMSE se alcanza en $k = 5$ así que usaremos ese valor.

Ejemplo: vinos.csv

Si aplicamos el modelo lineal que hemos obtenido al conjunto de test obtenemos:



Con valor de MSE del 0.32.

Conclusiones

Los valores predichos son bastante buenos porque en media se falla en menos de medio grado de alcohol para cada muestra del conjunto de test.

Aún así, observamos que los posibles valores resultados están acotados en el intervalo $(9, 14)$ por lo que los posibles errores cometidos van a estar también acotados.

Bibliografía

- Attouch M. K. & Bouabça W. (2012). *The k-nearest neighbors estimation of the conditional mode for functional data*. Rev.Roumaine Math. pures appl.**58**, *4*, 393-415.
- Chen, Y-C (2018) *Lecture 7: Density Estimation: k-Nearest Neighbour and Basis Approach* STAT 425: Introduction to Nonparametric Statistics
- Dasgupta, S. (2012). *Consistency of Nearest Neighbor Classification under Selective Sampling*. Workshop and Conference Proceedings, **23**,18.1-18.15.
- Fehrmann, L. & Kleinn, C. (2005). *A k-nearest neighbor approach for estimation of single-free biomass*. Proceedings of the Seventh Annual Forest Inventory and Analysis Symposium.

Bibliografía

- Guyader, A. & Hengartner, N. (2013). *On the Mutual Nearest Neighbors Estimate in Regression*. Journal of Machine Learning Research, **14**, 2361-2376.
- Györfi, L. *A distribution-free theory of nonparametric regression*. Springter. 86-96.
- Härdle, W. (1994) *Applied nonoparametric Regression*. Cambridge University.
- Mack, Y. P. & Rosenblatt, M. *Multivariate k-Nearest Neighbour Density Estimates* J. Multivariate Anal. **9** (1979), 1-15.
- Magnussen, S., McRoberts R. E. & Tomppo, E. O. (2009). *Model-based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories*. Remote Sensing of Environment 113, 476-488.

Bibliografía

- Lai, S. L. (1977). *Large sample properties of k-nearest neighbor precedures*, Ph.d. dissertation, *Dept. Mathematics, UCLA*. Los Ángeles.
- Lian, H. (2011). *Convergence of functional k-nearest neighbor regression estimate with funtional responses*. Electronic Journal of Statistics, **5**,31-39.
- Orava, J. (2011). *k-Nearest neighbour kernel density estimation, the choice of optimal k*. Trata Mt. Math., **50**, 39-50.
- Ruppert, D. & Wand, M. P. (1994). *Multivariate locally weighted least squares regression*. The Annals of Statistics. Vol.22, 1346-1370.
- Shao, J. (2009). *Nonparametric Variance Estimation for Nearest Neighbor Imputation*. Journal of Official Statistics, **25**, 55-62.

Bibliografía

- Stone, C. J. (1977). *Consistent nonparametric regression (with discussion)*. Annals of Statistics, **5**, 595-645.
- *https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/*
- *https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009*
- *https://rpubs.com/joser/caret*