Universidad de Buenos Aires
Facultad de Ingenieria
Año 2018 - 2er Cuatrimestre

Aprendizaje Estadístico, Teoría y aplicación

Trabajo Práctico Final

Sbruzzi, José Ignacio - Ingeniería Informática #97452
jose.sbru@gmail.com

Índice

1. Idea - introducción teórica

El objetivo es comprobar empíricamente el teorema 5.2 del Gyorfy:

Theorem 5.2: For a kernel estimate with a naive kernel assume that

$$Var(Y|X = x) \leq \sigma^2, x \in \mathbb{R}^d$$

and

$$|m(x) - m(z)| \leq ||x - z||, x, z \in \mathbb{R}^d$$

and X has a compact support S^*. Then

$$\mathbb{E}||m_n - m||^2 \leq \hat{c}\frac{\sigma^2 + sup_{z \in S^*} |m(z)|^2}{n \cdot h_n^d} + C^2 h_n^2$$

where $\hat{c}$ depends only on the diameter of S^* and on d, thus for

$$h_n = c'\left(\frac{\sigma^2 + sup_{z \in S^*} |m(z)|^2}{C^2}\right)^{1/(d+2)} n^{\left(-\frac{1}{d+2}\right)}$$

we have

$$\mathbb{E}||m_n - m||^2 \leq c''\left(\sigma^2 + \sup_{z \in S^*} |m(z)|^2\right)^{2/(d+2)} C^{2d/(d+2)} n^{-2/(d+2)}$$

Para esto, se acota el problema a la siguiente situación:

- $x_i \in D$
- $|m(z)| \leq 1$ para todo $z \in D$
- El ruido agregado a $m(z)$ para generar los pares x_i, y_i es una normal estándard, con lo cual $Var(Y|X = x) = 1$, es decir, $\sigma = 1$

Siendo $D = [-1, 1]^d$

Así, queda acotado también C. De esta forma, la última ecuación del teorema puede escribirse como:

$$\mathbb{E}||m_n - m||^2 \leq c''(1 + 1)^{2/(d+2)} C^{2d/(d+2)} n^{-d/(d+2)}$$

Podemos hacer algo similar con la primera conclusión del teorema:

$$\mathbb{E}||m_n - m||^2 \leq \hat{c}\frac{1 + 1}{n \cdot h_n^d} + C^2 h_n^2$$

2. Primera estrategia

2.1. Estimación de $\mathbb{E}||m_n - m||^2$

A continuación se explican los pasos que usa el programa para estimar este valor para determinados n, d y h_n.

1. generar una función $m(\cdot)$ con $-1 \leq m(x) \leq 1$ para todo $x \in D$.
2. generar una función $s(x)$ con las mismas características
3. Generar un conjunto P de n pares (x_i, y_i) tales que $y_i = m(x_i) + S$, donde S tiene una distribución normal centrada en 0 y con una varianza $|s(x_i)| \leq 1$. Los puntos x_i pertenecen a D, es decir, tienen d dimensiones.
4. A partir de este conjunto P de pares, se genera una estimación de la regresión, m_n, usando un naive kernel y el h_n correspondiente.
5. Teniendo $m(x)$ y $m_n(x)$ definidos para todo $x \in D$, se utiliza la librería de python mcint para integrar $(m(x) - m_n(x))^2$ sobre todo D. mcint utiliza técnicas montecarlo para estimar la integral, ya que para d dimensiones la integral es difícil de calcular numericamente (es decir, tarda demasiado). Así se obtiene un $||m - m_n||^2$.
6. Se repite este procedimiento para una cantidad de $m(\cdot)$, $s(\cdot)$ y $m_n(\cdot)$ generadas al azar (en la mayoría de los casos se hicieron 300 experimentos para cada n y d, en otros casos se hicieron 100).
7. Se promedian los $||m - m_n||^2$ para obtener una estimación de la esperanza.

Así, se obtiene la función $encontrarEError(n, d, h_n)$.

2.2. Verificación del teorema

La idea inicial era verificar que al variar n y mantener fijo d y h_n, se cumpliría que existe una cota de la forma

$$c(n^{-k})$$

que cumpla:

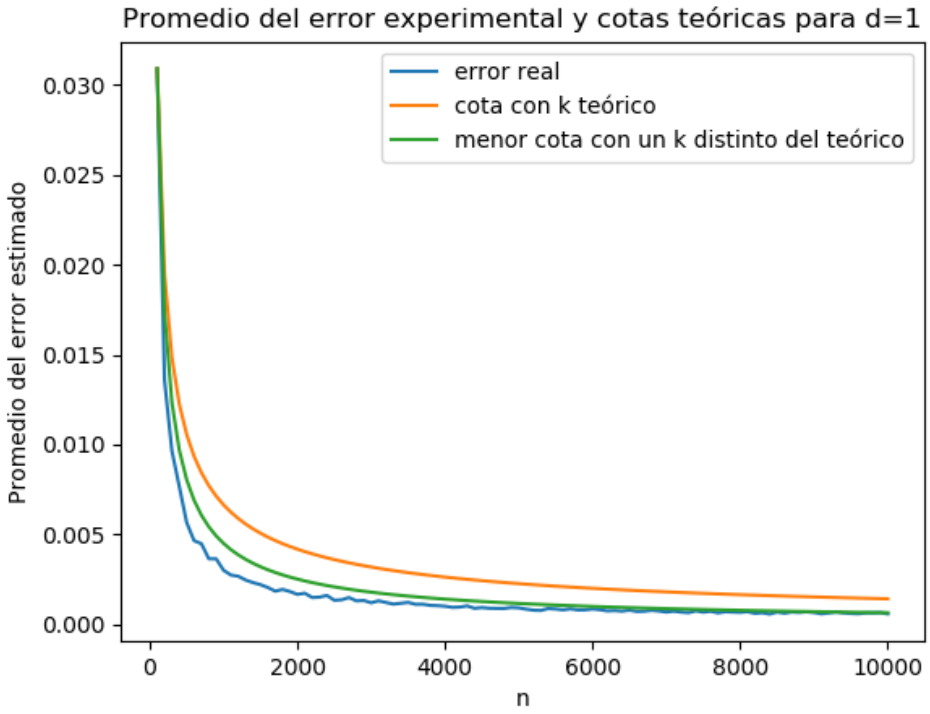
- $c(n^{-k})$ es mayor que todas las estimaciones $encontrarEError(n)$ (d y h_n son fijos)

- Los c y k elegidos deben ser tales que minimicen $\sum_{i=1}^{n} c(n^{-k})$

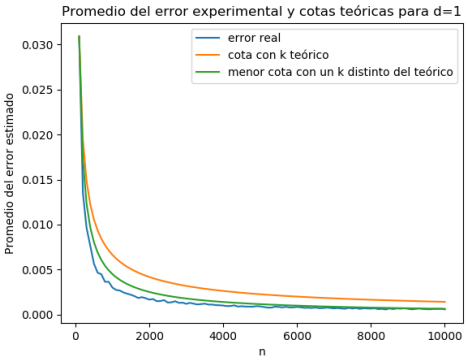
Así, la curva más ajustada a los datos (es decir, con c y k mejores que los que propone el teorema), debería cumplir que k sea mejor al propuesto por el teorema (el teorema indica $k = 2/(d+2)$) para verificarlo.

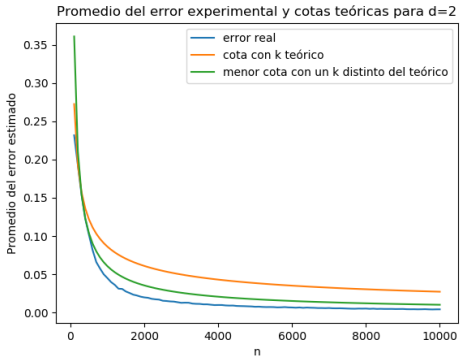
También se analizó la curva que cumple $k = 2/(d+2)$. En este caso simplemente se agregó esta condición sobre k y se buscó sólo el c que cumpla las condiiciones listadas arriba.

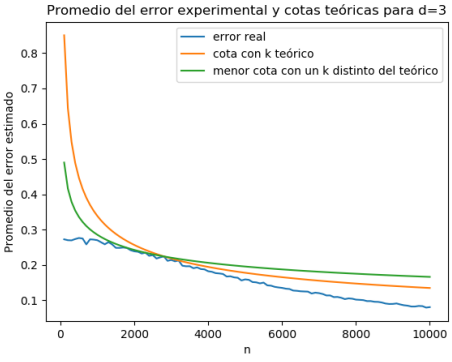
Esta prueba se repitió para $h_n = 0{,}5$ y $h_n = 0{,}1$.

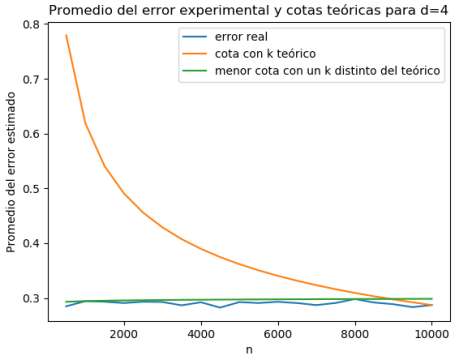


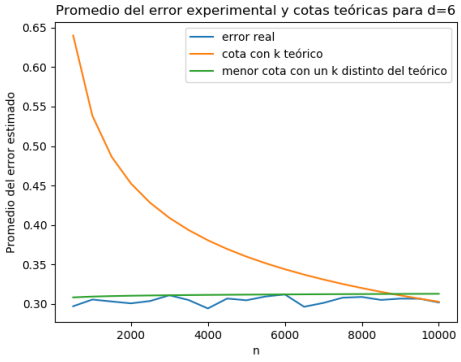
3. Primeros resultados: $h_n = 0{,}1$

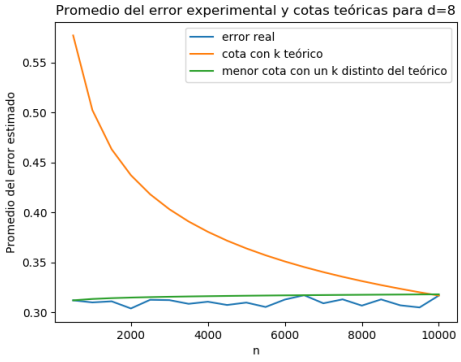


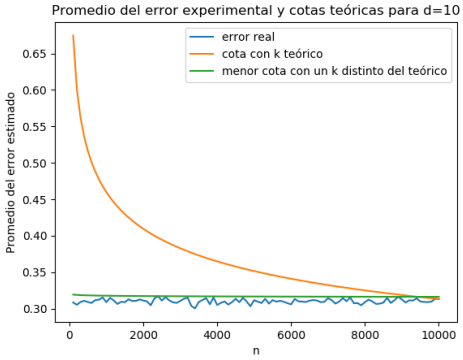


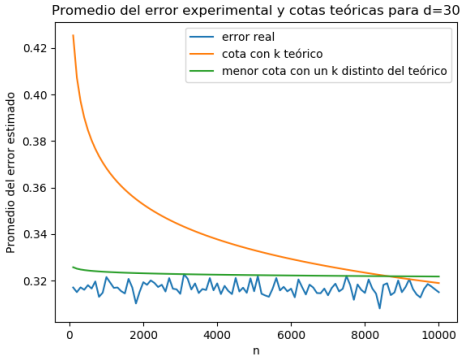


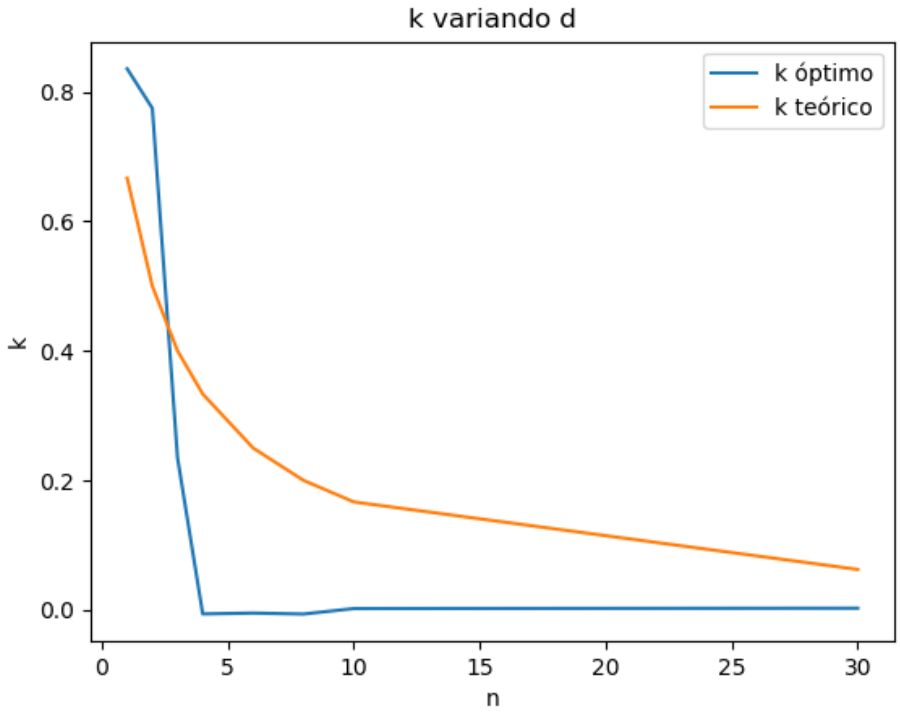


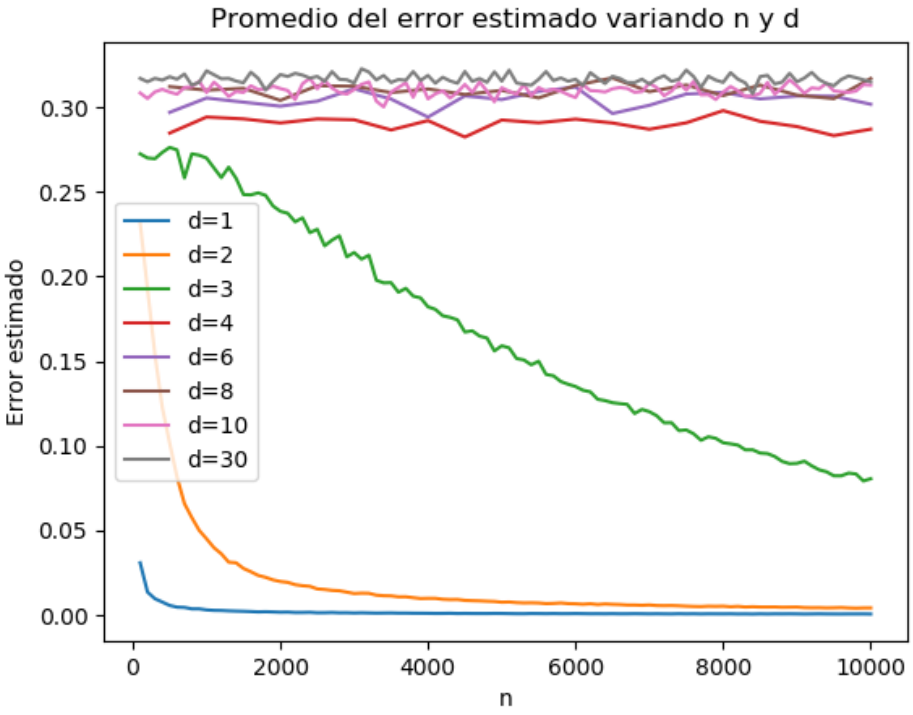




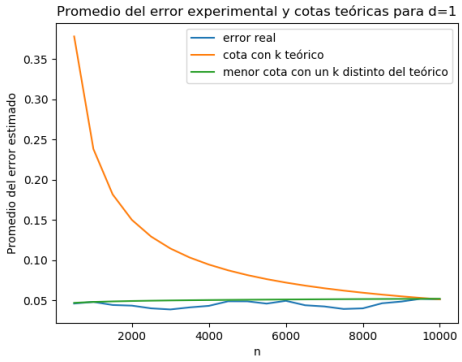

Promedio del error experimental y cotas teóricas para d=10
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

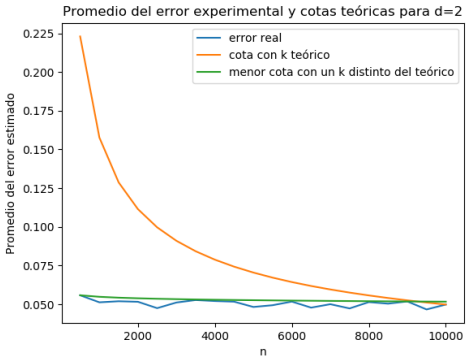

Promedio del error experimental y cotas teóricas para d=30
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
n

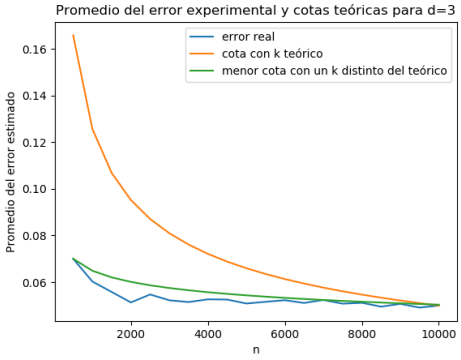

k variando d
k óptimo
k teórico
k
n

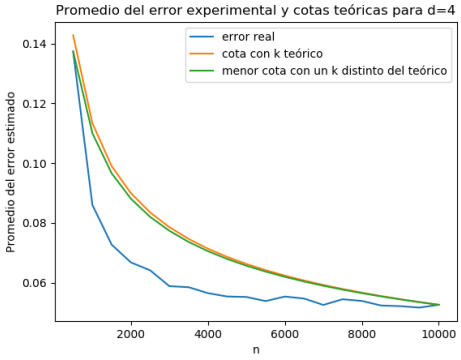

Promedio del error estimado variando n y d
Error estimado
n
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2000
4000
6000
8000
10000
d=1
d=2
d=3
d=4
d=6
d=8
d=10
d=30

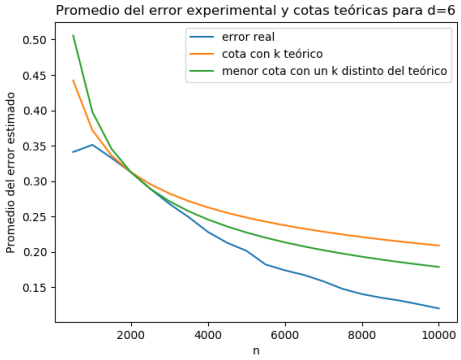
4. Primeros resultados: $h_n = 0{,}5$

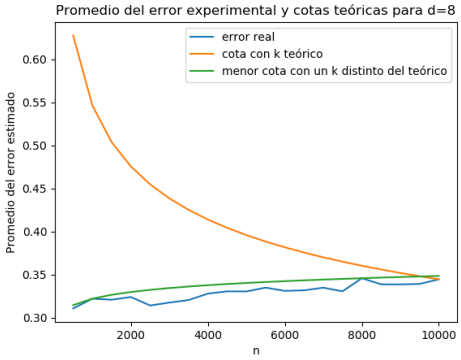


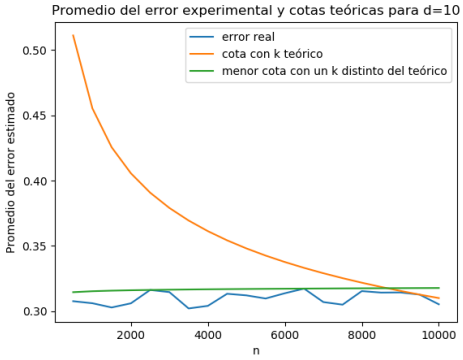


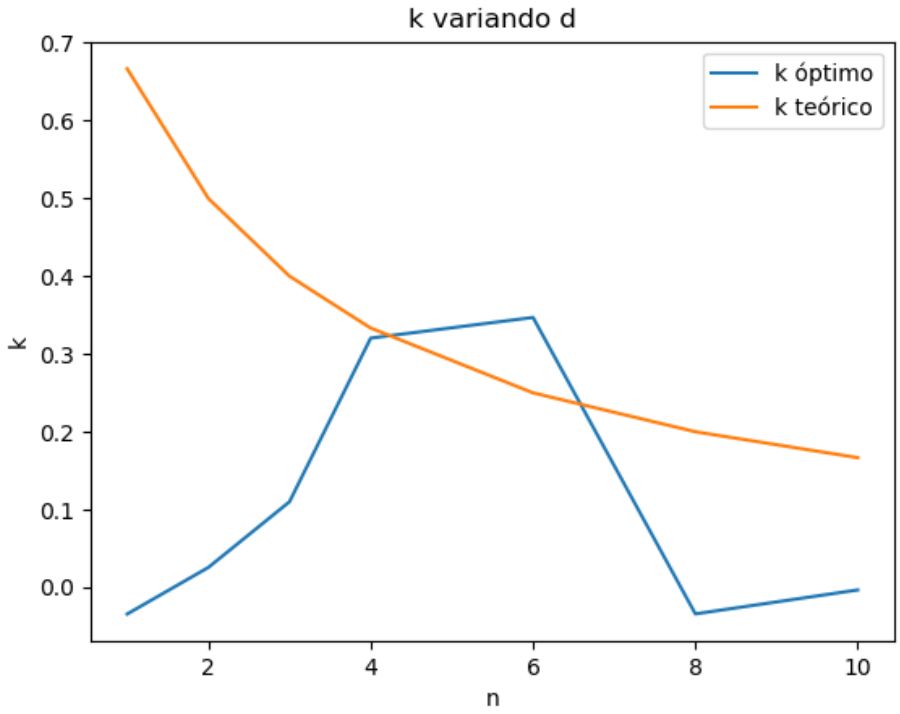


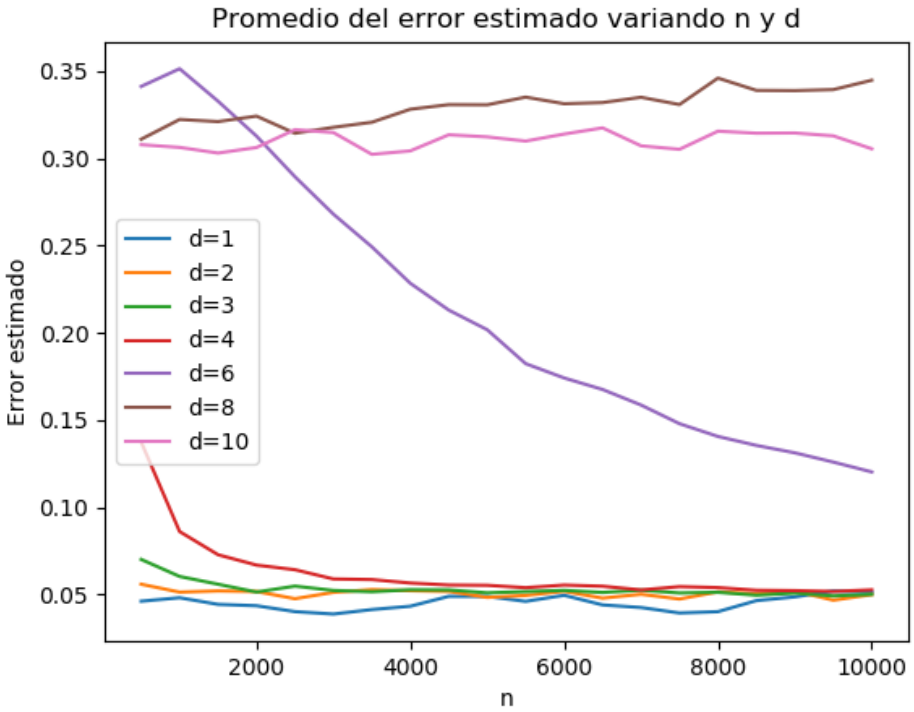







Promedio del error experimental y cotas teóricas para d=10
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
0.50
0.45
0.40
0.35
0.30
2000
4000
6000
8000
10000
n


k variando d
k óptimo
k teórico
k
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2
4
6
8
10
n



5. Conclusiones para h_n constante

Para $h_n = 0{,}1$ se obtuvieron resultados buenos (que verifican) para $d = 1$, $d = 2$ y $d = 3$, pero para d superiores no se logró la verificación.

Esto es razonable cuando se tienen en cuenta las reglas usadas en la práctica çomún"de machine learning: al aumentar la cantidad de dimensiones y no subsanar esto con más datos se tiene underfitting.

En este caso en particular también es importante la maldición de la dimensionalidad: a medida que crece d, la cantidad de puntos a una distancia fija h_n cae (en realidad, el mismísimo significado de la distancia es el que se pierde: todos los puntos tienden a estar a una distancia muy similar unos de otros).

Aunque las reglas prácticas del machine learningz la maldición de la dimensionalidad explican los resultados, estos contradicen al teorema que se busca corroborar empiricamente (es decir, la interpretación que se había hecho del mismo).

Con el objetivo de observar el comportamiento del algoritmo en dimensiones altas, se utilizó $h_n = 0{,}5$ para realizar una nueva prueba.

Los resultados de esta prueba son muy importantes: para dimensiones

"medianas"($d = 4$, $d = 6$), con $h_n = 0{,}5$ se logra lo que no se puede con $h_n = 0{,}1$: se tiene una mejora gradual y se corrobora el teorema para $d = 6$.

El problema es que fue ignorada la condición sobre h_n que requiere la segunda conclusión del teorema:

$$h_n = c' \left(\frac{\sigma^2 + sup_{z \in S^*} |m(z)|^2}{C^2} \right)^{1/(d+2)} n^{\left(-\frac{1}{d+2}\right)}$$

6. Segunda estrategia

Es imposible fijar

$$h_n = c' \left(\frac{\sigma^2 + sup_{z \in S^*} |m(z)|^2}{C^2} \right)^{1/(d+2)} n^{\left(-\frac{1}{d+2}\right)}$$

ya que para eso sería necesario conocer c'.

Entonces se intenta corroborar el teorema 5.1, que establece:

> **Theorem 5.1:** Assume that there are balls $S_{0,r}$ of radius r and balls $S_{0,R}$ of radius R centered at the origin ($0 < r \leq R$), and constant $b > 0$ such that
>
> $$\mathbb{1}\{x \in S_{0,R}\} \geq K(x) \geq b\mathbb{1}\{x \in S_{0,r}\}$$
>
> (boxed kernel), and consider the kernel estimate m_n if $h_n \to 0$ and $nh_n^d \to \infty$, then the kernel estimate is weakly universally consistent.

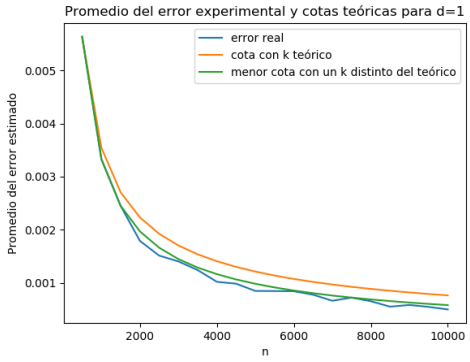
El kernel naive es un boxed kernel, por lo tanto, si se cumplen las condiciones sobre h_n que establece este teorema, se obtiene la consistencia debil.

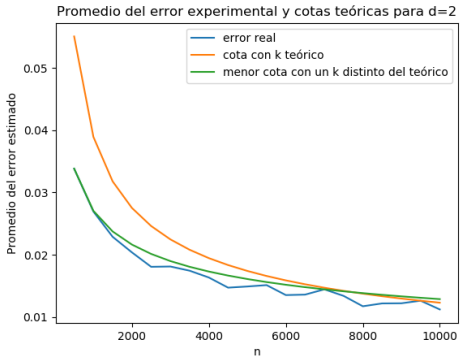
Así, se llevaron adelante dos pruebas: una con h_n dependiendo de n y d, y otra en la cual depende sólo de d. Así, para la primera, se usó

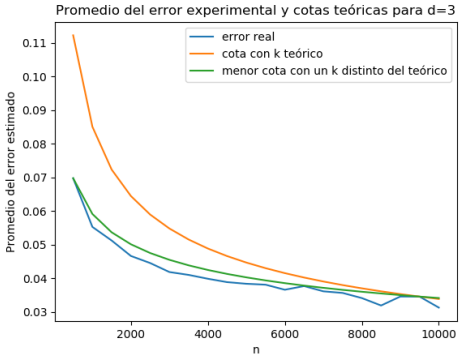
$$0{,}8548(n^{(-1/(4{,}054 \cdot d))})$$

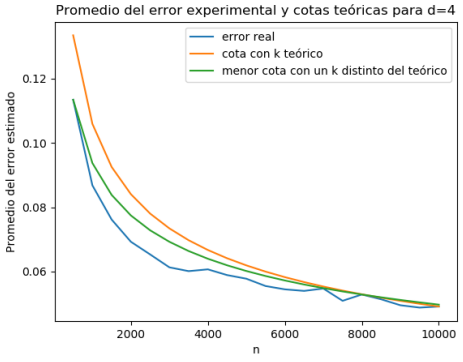
, lo cual cumple las condiciones y además cumple que h_n es aproximadamente 0,1 cuando $d = 1$ y $n = 8000$, y aproximadamente 0,5 cuando $d = 4$ y $n = 8000$. Para la segunda corrida de pruebas se utilizó $h_n = 10^{-1/d}$, elegido con el mismo criterio.

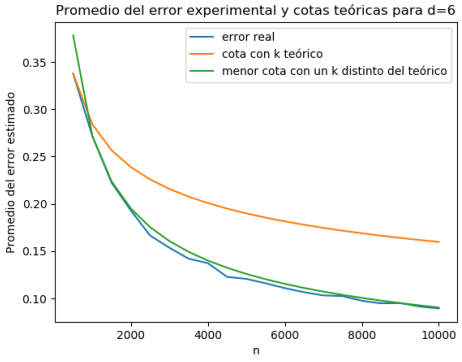
7. Resultados para h_n variando con d y n

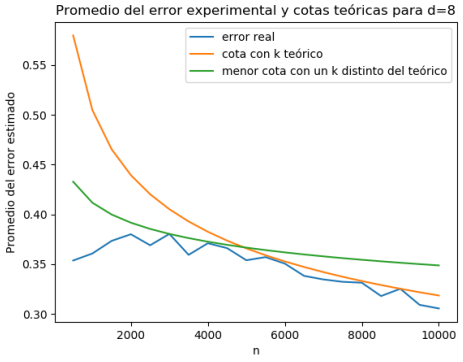


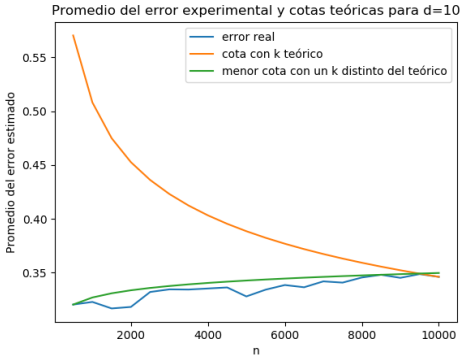


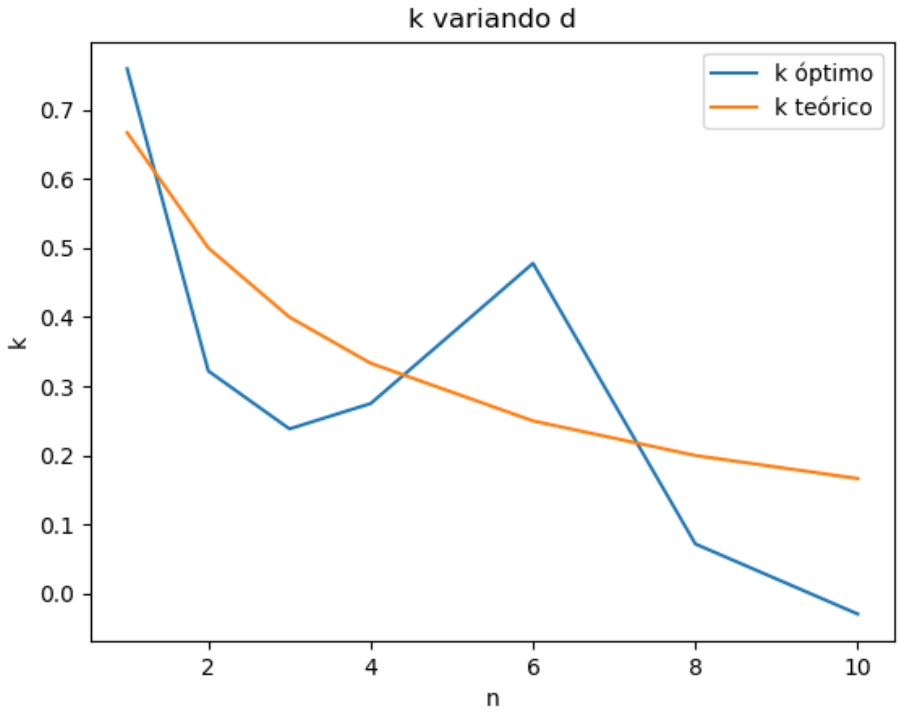


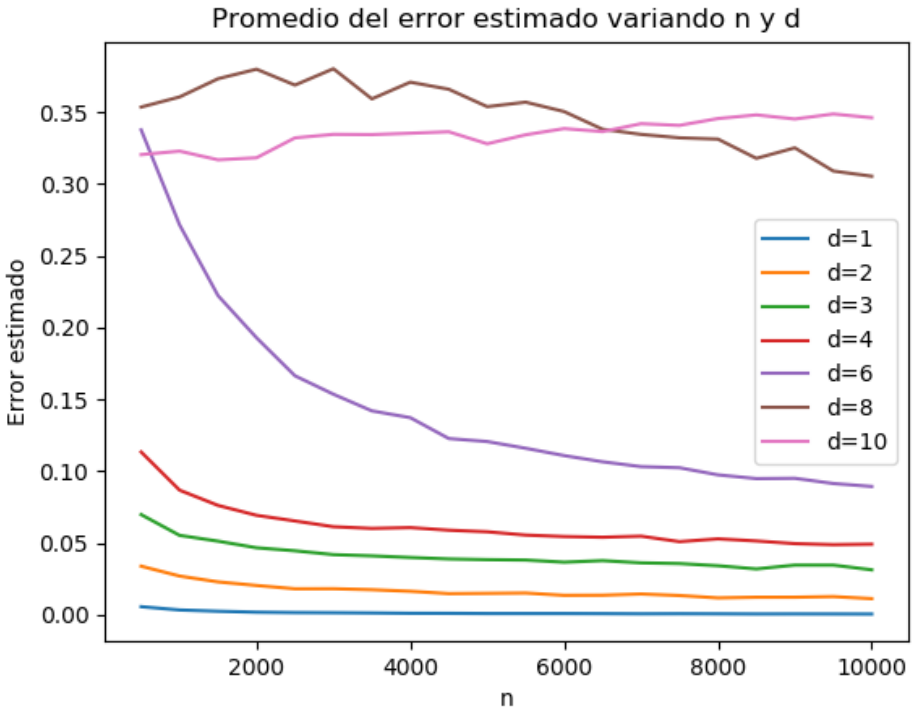




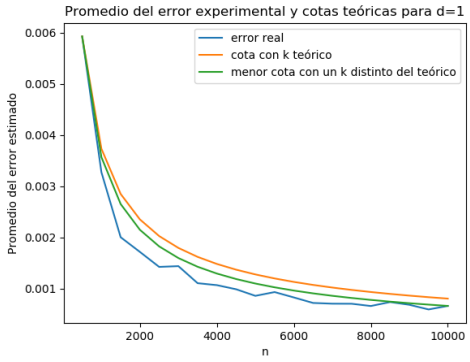


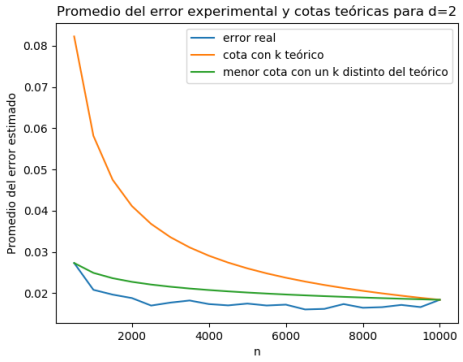

Promedio del error experimental y cotas teóricas para d=10
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
0.55
0.50
0.45
0.40
0.35
2000
4000
6000
8000
10000
n

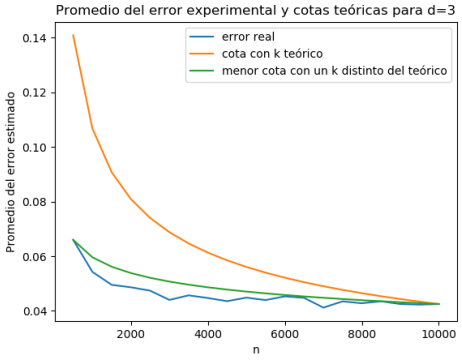

k variando d
k óptimo
k teórico
k
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2
4
6
8
10
n

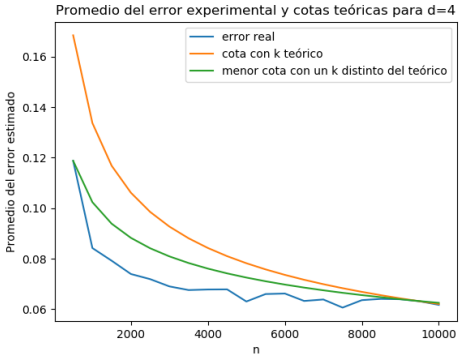

Promedio del error estimado variando n y d
Error estimado
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2000
4000
6000
8000
10000
n
d=1
d=2
d=3
d=4
d=6
d=8
d=10

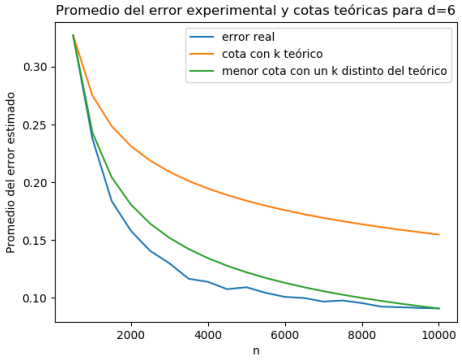
8. Resultados para h_n variando sólo con d

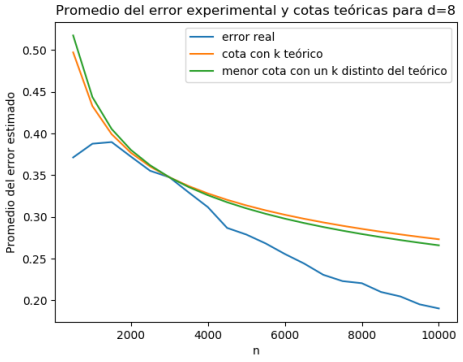


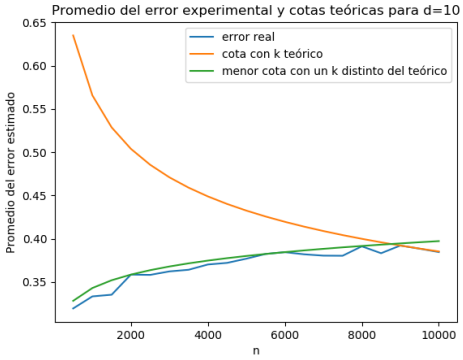


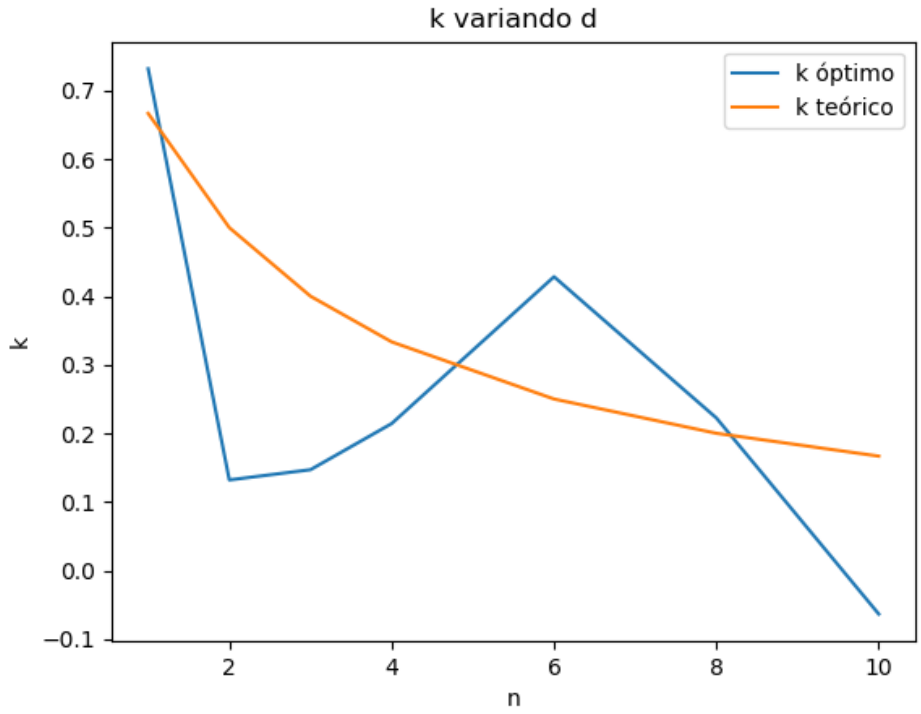


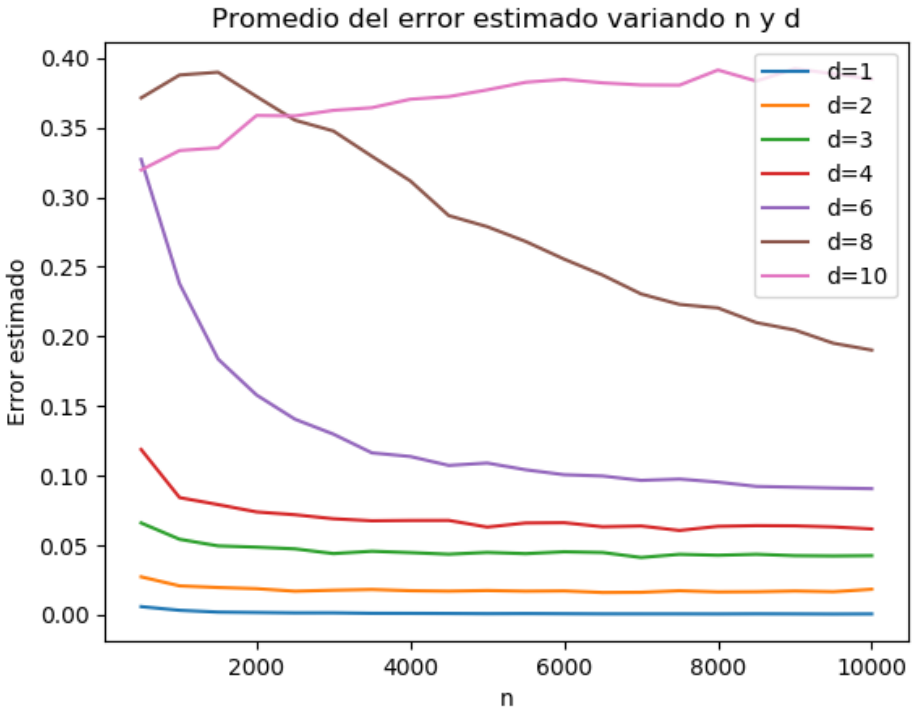







Promedio del error experimental y cotas teóricas para d=10
error real
cota con k teórico
menor cota con un k distinto del teórico
Promedio del error estimado
0.65
0.60
0.55
0.50
0.45
0.40
0.35
2000
4000
6000
8000
10000
n


k variando d
k óptimo
k teórico
k
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
2
4
6
8
10
n



9. Conclusiones para h_n variable

En caso de h_n variable, se obtuvieron mejores resultados para todas las curvas: hubo aprendizaje para d entre 1 y 6, lo cual era de esperar debido al criterio utilizado para elegir las constantes, discutido previamente; y también hubo aprendizaje para $d = 8$, lo cual no se había observado para $h_n = 0{,}1$ ni $h_n = 0{,}5$. Esto es un indício fuerte para opinar que es posible corroborar el teorema 5.1.

El aprendizaje para $d = 8$ es muy notorio para h_n dependiente sólo de d, y menos notorio para h_n dependiente de n, lo cual podría resultar contradictorio al teorema.

El hecho de que para $d = 10$ no se dé el aprendizaje podría deberse a dos motivos:

1. Las constantes elegidas para las funciones $h_n(n, d)$ y $h_n(d)$ son incorrectas.

2. Las formas generales elegidas para $h_n(n, d)$ y $h_n(d)$ son incorrectas.

Otra posibilidad es que el teorema sea imposible de corroborar en altas dimensiones debido a la maldición de la dimensionalidad: que el aprendizaje sea demasiado lento como para observarlo.