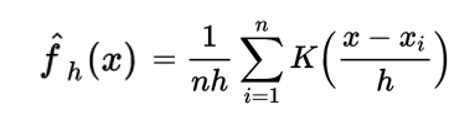
Trabajo Práctico 2 – Kernel Density Estimation

# Determinación de las funciones de Densidad de Probabilidad

Se determinaron las pdf de los clientes seguros f(z|S), de los clientes riesgosos f(z|R) y de todos f(z) con el método de Kernels. Los Kernels utilizados fueron el Lineal, el Gaussiano y el de Epanechnikov.



Para las condicionales sobre los clientes seguros o riesgosos, se filtra según corresponda en la muestra. Una vez determinados las pdf se procedió a determinar las funciones objetivo para cada uno de los dos casos que se desean analizar:

1. El banco desea maximizar la cantidad de préstamos a otorgar, mientras que la

tasa de mora en su cartera no supere el 1.5% (medida como proporción de la

cantidad de préstamos otorgados)

1. El banco desea maximizar el beneficio económico total. Para ello, se considera que si un préstamo llega a término se realiza un beneficio en valor presente de 1$, mientras que por cada préstamo que cae en default se genera una pérdida de $10.

En el caso a) entonces se desea encontrar el máximo z\* tal que P(R | z < z\*) < 1,5%. Como esta función es creciente en z\*, la optimización se transforma en encontrar el z\* que hace P(R | z < z\*) =1,5%. Ahora bien, por el teorema de Bayes, resulta:

P (R | z < z\*) = P( z < z\* | R) x P(R) / P(z < z\*)

Con las pdf calculadas al inicio podemos integrarlas y obtener F(z\*|R) = P( z < z\* | R) y F(z\*) = P(z < z\*) , con lo que quedan definidas las probabilidades de la fórmula anterior. Adicionalmente se sabe que la proporción de clientes de riesgo a clientes seguros es de 0,2. Con lo que resulta P(R)= 1/6. Finalmente, se encuentra z\* como el valor que hace que dicha fórmula sea igual a 1,5%, mediante inspección de un array donde ya se obtuvo el cociente de las probabilidades.

En el caso b) se desea encontrar el z\* que maximiza el ingreso:

I = [ 1 x P (S | z < z\*) – 10 x P (R | z < z\*)] x P (z < z\*) x N

De nuevo usando Bayes se obtiene

I = [ P (z < z\* | S) – 2 x P (z < z\* | R)] x 5 x N/6

El problema para hallar z\* consiste en maximizar dicha función, pero nuevamente ya conocemos todos los términos de dicha función integrando las pdf, así que simplemente se reduce a encontrar el máximo de un array ya calculado.

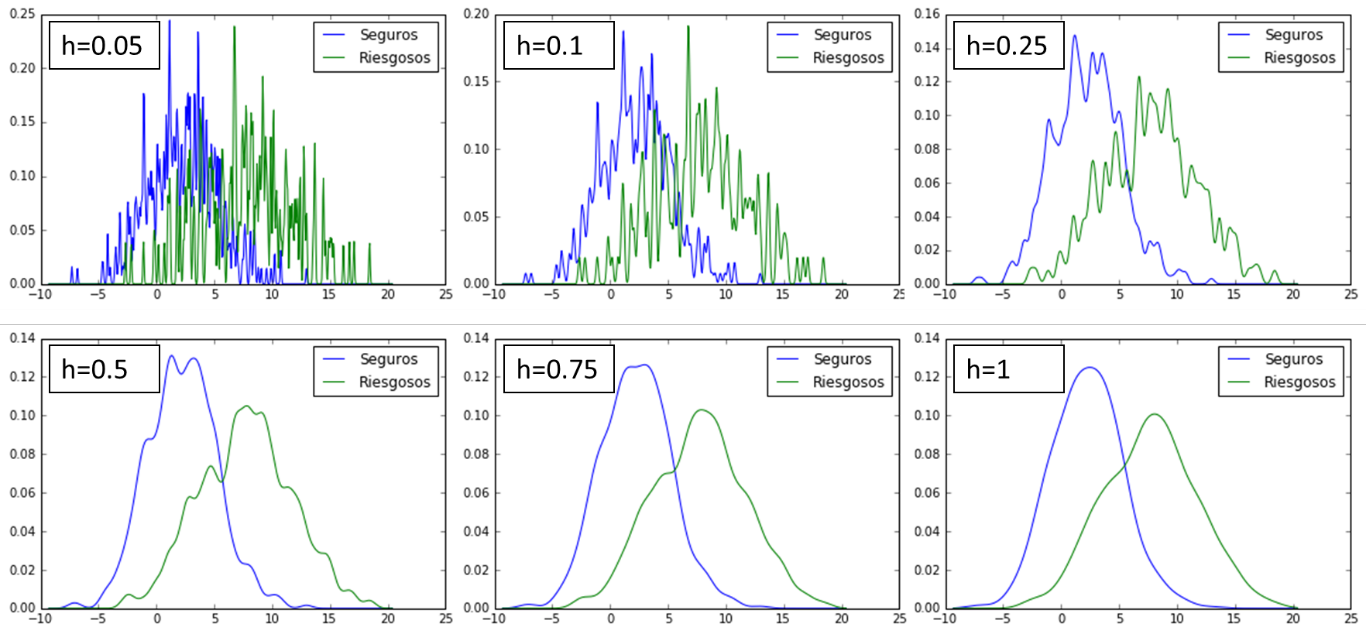
# Resultados

El resultado va a depender del ancho de banda h que se elija para el cálculo de las pdf, por lo que se resumen los valores obtenidos para el z\* en función de h en la siguiente tabla.

|  |  |  |  |
| --- | --- | --- | --- |
| **z\* caso (a)** | | | |
| **h** | **Linear** | **Gaussian** | **Epa** |
| 0,05 | -2,70 | -2,70 | -2,70 |
| 0,1 | -2,70 | -2,70 | -2,70 |
| 0,25 | -2,70 | -2,70 | -2,65 |
| 0,5 | -2,65 | -2,80 | -2,84 |
| 1 | -2,75 | -3,25 | -3,40 |
| 1,5 | -2,90 | -4,00 | -4,20 |
| 2 | -3,10 | -5,20 | -5,15 |
| 3 | -3,65 | -9,30 | -7,90 |
|  |  |  |  |
| **z\* caso (b)** | | | |
| **h** | **Linear** | **Gaussian** | **Epa** |
| 0,05 | 3,70 | 3,70 | 3,70 |
| 0,1 | 3,70 | 3,70 | 3,70 |
| 0,25 | 3,70 | 3,70 | 3,90 |
| 0,5 | 3,70 | 3,80 | 3,80 |
| 1 | 3,80 | 3,50 | 3,45 |
| 1,5 | 3,65 | 3,40 | 3,45 |
| 2 | 3,55 | 3,20 | 3,15 |
| 3 | 3,55 | 2,55 | 2,35 |

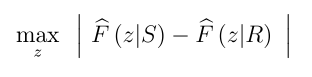
Se puede observar que en ambos casos, los diferentes Kernels muestran resultados aproximadamente equivalentes. Esto en particular ocurre cuando h está por debajo de 1. Para simplificar el análisis, se grafican los resultados en función de h.

Ahora bien, la determinación de h juega un rol importante, ya que en función de dicho parámetro vemos que la determinación del z\* puede variar. Para evaluar la bondad de la determinación del h, en primera instancia se revisaron las pdf resultantes. A continuación, se observan los gráficos para el caso del Kernel Gaussiano:



Resulta claro que cuando h es muy pequeño el ruido domina la distribución, y a medida que aumenta el ancho de banda, dicho ruido se va atenuando. Cuando se llega a h=1 las curvas ya están suavizadas.

Sumado a esto, es posible determinar la bondad del modelo de scoring para separar los clientes buenos de los malos mediante el índice de Kolmogorov- Smirnov (KS), que se obtiene como:



KS mide el solape entre las distribuciones de seguros y riesgosos, siendo KS=1 el caso ideal donde no hay solape alguno, y las curvas de F tienen un rango de z donde F(z|S)=1 y F(z|R)=0. El peor caso, es cuando KS=0, que básicamente indica que F es la misma para los seguros y los riesgosos, y por lo tanto el modelo de scoring es incapaz de separar a los clientes.

En los casos intermedios que son los que nos interesan. Aquí supongamos que z es el score donde se da el máximo que define KS. Vemos que F(z|R) nos da la probabilidad de que un malo sea admitido y (1 - F(z|S)) la probabilidad que un bueno no sea admitido. A su vez, por definición KS + F(z|R) + (1 - F(z|S)) = 1. Entonces, conceptualmente a medida que KS se hace más grande, se reduce la probabilidad de rechazar un bueno y de aceptar un malo.

Dicho esto, los resultados de KS en función de h se muestran en la siguiente tabla y gráfico.

|  |  |  |  |
| --- | --- | --- | --- |
| **KS** | | | |
| **h** | **Linear** | **Gaussian** | **Epa** |
| 0,05 | 0,59 | 0,59 | 0,59 |
| 0,1 | 0,59 | 0,59 | 0,59 |
| 0,25 | 0,59 | 0,58 | 0,58 |
| 0,5 | 0,58 | 0,57 | 0,57 |
| 1 | 0,58 | 0,56 | 0,56 |
| 1,5 | 0,57 | 0,53 | 0,53 |
| 2 | 0,57 | 0,51 | 0,50 |
| 3 | 0,55 | 0,45 | 0,44 |

En el caso Lineal, el decaimiento es más atenuado y resulta razonable utilizar h=1 sin sacrificar casi nada de KS. Para el caso Gaussiano y el cuadrático, pasar de h= 0,5 a h=1 ya implica una reducción de KS, pero que sin embargo se justifica por lo visto de las pdf que se obtienen con cada nivel de h.

Adicionalmente, maximizar KS esconde una trampa en el hecho que no tiene en cuenta la proporción relativa entre buenos y malos. En el caso que la proporción de malos sea mucho menor que la de buenos, no tiene sentido prestarle atención a la probabilidad de aceptar uno malo, por ejemplo.

Como conclusión, se encontraron los z\* que optimizan los dos problemas para distintos valores del ancho de banda. Como resultado final, de tener que elegir un valor de z\* para el caso a) tomaría z\*=-2,8, y para el caso b) tomaría z\*=3,8 que corresponde al Lineal con h=1 y Gaussiano y Epa con h=0,5.