Reconocimiento de Patrones

Practico 2

Autor:
Agustín Navcevich 4.646.135-9

Mail:
agus-navce52@hotmail.com

Instituto de Ingeniería Eléctrica

Facultad de Ingeniería

Universidad de la República

9 de octubre de 2017

Ejercicio 1

1.1

Para mostrar que esto primero se vera el caso donde $P(w_{max}|x) = P(w_i|x)\ \forall i = 1, ..., c$.
En este caso es trivial que el maximo es $\frac{1}{c}$ ya que toda $P(w_i|x) = \frac{1}{c}$, por lo que es trivial ver que se cumple la desigualdad.

Ahora para los demas casos, primero sabemos que $\sum_{i=1}^{c} P(w_i|x) = 1$.Tambien se sabe a partir de esto que si se toma como el caso anterior donde cada $P(w_i|x) = \frac{1}{c}$, entonces la única forma de poder cambiar algún valor para las probabilidades es que si alguna aumenta, por lo menos alguna de las otras probabilidades tiene que disminuir dado que se esta limitado porque la suma de todas las probabilidades sea igual a uno.
Entonces visto esto, si aumento una de las probabilidades por encima de $\frac{1}{c}$ sabemos que todas las demas van a ser menores que esta primera que pasa a ser $P(w_{max}|x)$.
Por lo mencionado anteriormente, las consiguientes probabilidades tienen un valor que es menor o igual a $\frac{1}{c}$ por la disminución que se debe dar dado que $P(w_{max}|x)$ aumento. Entonces, tenemos que siempre $P(w_{max}|x) \geq \frac{1}{c}$ para cualquier eleccion de $P(w_i|x)$.

1.2

Si se define la probabilidad de error como sigue:

$$P(error) = \int P(error|x)p(x)dx$$

Entonces para la regla de decisión $P(w_{max}|x) \geq P(w_i|x)\ \forall i = 1, ..., c$ tenemos que:

$$P(error|x) = min\left[P(w_1|x), ..., P(w_c|x)\right]$$

Por lo cual tenemos:

$$P(error) = \int P(w_{min}|x)p(x)dx$$

que es igual a:

$$P(error) = 1 - \int P(w_{max}|x)p(x)dx$$

1.3

Imponiendo la desigualdad de ejercicio uno a la ecuacion de la parte 2 tenemos:

$$P(error) \leq 1 - \int \frac{1}{c}p(x)dx$$

y como integra a uno $p(x)$ ahi obtenemos el resultado:

$$P(error) \leq 1 - \frac{1}{c} = \frac{c-1}{c}$$

1.4

La situación donde se cumple que $P(error) = \frac{c-1}{c}$ es cuando todas las probabilidades son iguales que es cuando la desigualdad se convierte en una igualdad para la ecuación del problema de la parte 1.

Ejercicio 2

2.1

Sea $f_0(x) = N(0, \sigma^2)$ y $f_1(x) = N(\mu, \sigma^2)$, entonces con γ definido tenemos el plano dividió en dos para lo que se llamara R_1 la región de decisión por H_1 y R_0 a la de H_0, como se ve en la figura 1.

A partir de esto usamos la definición de la probabilidad de detección: y llegamos a que:

$$P_D = \int_{R_1} f_1(x)dx$$

También usando la definición de probabilidad de falsas alarmas se obtiene:

$$P_{FA} = \int_{R_1} f_0(x)dx$$

Y como se esta trabajando con gaussianas entonces obtenemos en cada una de las definiciones la cola de la gaussiana para sus distintas distribuciones comenzando desde γ, entonces se tiene que:

$$P_D = Q(\frac{\gamma - \mu}{\sigma})$$

$$P_{FA} = Q(\frac{\gamma}{\sigma})$$

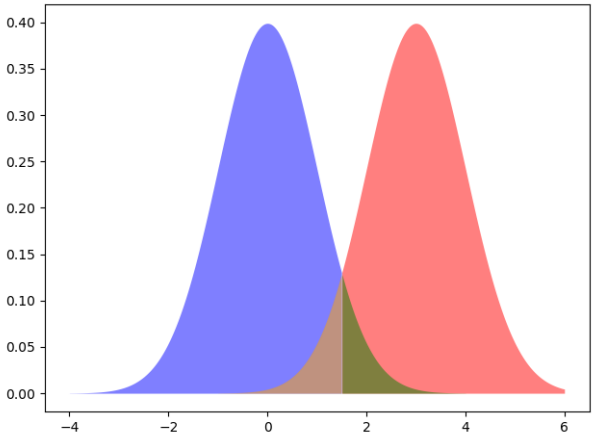


Figura 1: Curva ROC

Como se aprecia en la figura 1 en color verde tenemos la probabilidad de falsas alarmas, que es el área bajo la curva de $f_0(x)$ en R_1 y el área pintada de rosado mas la parte verde de la curva es la probabilidad de detección.

2.2

Para hallar P_D en función de P_{FA} primero se tiene que:

$$\gamma = \frac{Q^{-1}(P_{FA})}{\sigma}$$

Y sustituyendo en la expresión para P_D:

$$P_D(P_{FA}) = Q\left(\frac{\frac{Q^{-1}(P_{FA})}{\sigma} - \mu}{\sigma}\right)$$

La curva es una representación gráfica de la sensibilidad frente a la 1 menos la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Otra interpretación de este gráfico es la representación de la razón de verdaderos positivos frente a la razón o ratio de falsos positivos según se varía el umbral de discriminación (valor a partir del cual decidimos por H_0 o H_1).

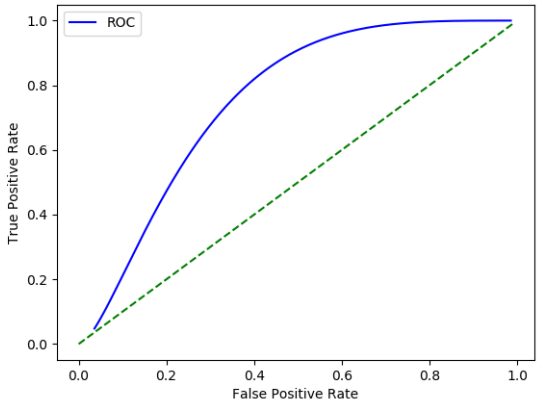


Figura 2: Curvas de la distribuciones

2.3

Para diseñar una regla de decisión con $P_{FA} \leq \alpha$ se utiliza el lema de Neyman-Pearson que dice que la regla que maximiza P_D bajo esta restricción es eligiendo η de forma que:

$$P_{FA} = \int_{x:\Delta(x)>\eta} f_0(x)dx = \alpha$$

Donde $\Delta(x)$ es lo que se utiliza para realizar la elección de umbral y es $\Delta(x) = \frac{f_1(x)}{f_0(x)}$.

Ejercicio 3

3.1

El algoritmo que sigue el perceptron es el siguiente:

- Se inicializan el vectro w de forma aleatoria
- Se realiza con todos los datos de entrada la evaluacion de la ecuacion del perceptron:

$$y = \begin{cases} 1 \quad si \quad w^T x > 0 \\ \\ 0 \quad si \quad otrocaso \end{cases}$$

- Si y es distinto de la etiqueta que traían los datos entonces se actualizan los pesos como $w = w + error * x$
- se repite hasta que todos los puntos de datos estén correctamente clasificados:

3.2

Se muestra que se pueden separar los patrones junto con la región de decisión en la figura 3:

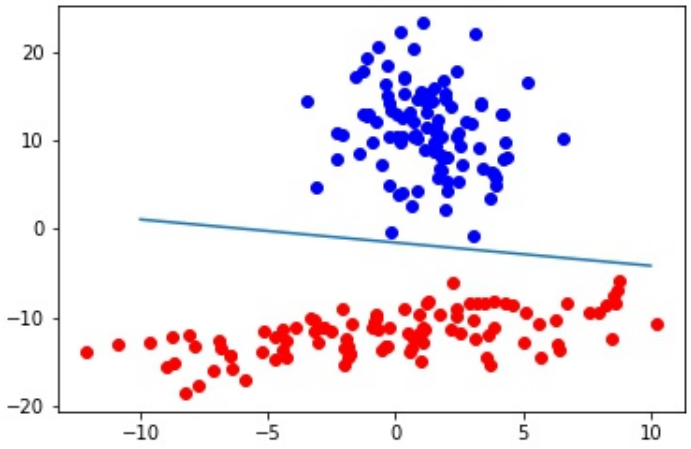


Figura 3: Datos separados por el perceptron junto con curva de decisión

3.3

Una descripción del problema XOR es que la salida debe estar encendida cuando cualquiera de las entradas está encendida, pero no cuando ambas están encendidas. Si vemos el problema con la ecuación del perceptron entonces cuatro desigualdades deben ser satisfechas para que la percepción resuelva este problema imponiendo los datos:

$$(0w_1) + (0w_2) < \theta \longrightarrow \theta > 0$$
$$(0w_1) + (1w_2) < \theta \longrightarrow \theta > w_2$$
$$(1w_1) + (0w_2) < \theta \longrightarrow \theta > w_1$$
$$(1w_1) + (1w_2) < \theta \longrightarrow \theta > w_1 + w_2$$

Sin embargo, esto obviamente no es posible ya que tanto w_1 como w_2 tendrían que ser mayores que cuando su suma $w_1 + w_2$ es menor que. Desde una perspectiva geométrica, el perceptrón intenta resolver los problemas de XOR usando una línea recta para separar las dos clases: las entradas marcadas con "0"deberían estar en un lado de la línea y las entradas marcadas con "1"deberían estar en el otro lado. Para las funciones XOR no es posible hacerlo porque no es una función linealmente separable.

Ejercicio 4

4.1

Los coeficientes optimos para el conjunto de datos fueron los siguientes:

$$w_1 = -0{,}00137293, w_2 = 0{,}00678947$$

4.2

En la figura 4 podemos ver el vector w de pesos junto a los datos:

4.3

En la figura 5 podemos observar como fueron clasificados los patrones y como fueron etiquetados.

Para este clasificador y estos datos se obtuvo una precisión del 98,5 % donde en la figura 5 podemos ver exactamente cuales fueron los puntos mal clasificados. Como sabemos, la recta de decisión es ortogonal a la dirección de el vector w. Observando la dirección de w antes de realizar la clasificación se podía observar que la recta que es ortogonal al vector w es una recta que deja los dos puntos que se clasificaron mal por debajo de la recta de decisión entonces debido a esto no separa los datos totalmente por lo cual se podía decir que los datos no iban a ser totalmente separables según este criterio.

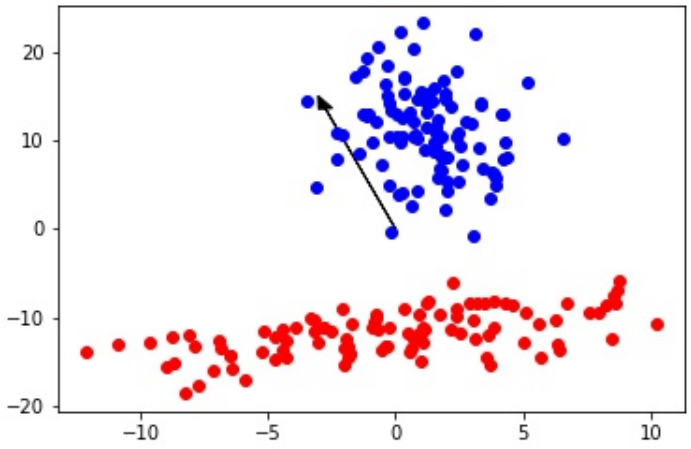


Figura 4: Datos separados por el perceptron junto con curva de decisión

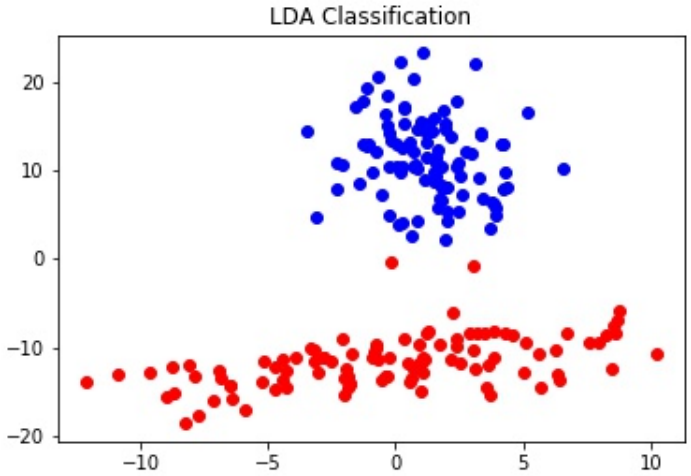


Figura 5: Datos separados por el perceptron junto con curva de decisión

Ejercicio 5

5.1

Tenemos la definición para la verosimilitud:

$$l(w|X,y) = logP(y|X)$$

Para regresión logística ya que tenemos que es una binomial entonces:

$$l(w|X,y) = log\left(\prod_{i=1}^{n} P(Y=y_i|X=x_i;W)^{y_i}(1-P(Y=y_i|X=x_i;W))^{1-y_i}\right)$$

$$l(w|X,y) = \sum_{i=1}^{n} y_i logP(Y=1|X=x_i;W) + (1-y_i)log(1-P(Y=1|X=x_i;W))$$

Entonces:

$$l(w|X,y) = \sum_{i=1}^{n} y_i log(\sigma(x_i,w)) + (1-y_i)log(1-\sigma(x_i,w))$$

Simplificando:

$$l(w|X,y) = \sum_{i=1}^{n} y_i log\left(\frac{\sigma(x_i,w)}{1-\sigma(x_i,w)}\right) + log(1-\sigma(x_i,w))$$

5.2

Aplicando la derivada a cada termino:

$$\frac{\partial l(w|X,y)}{\partial w_j} = \sum_{i=1}^{n} y_i \frac{\partial}{\partial w_j} log\left(\frac{\sigma(x_i,w)}{1-\sigma(x_i,w)}\right) + \frac{\partial}{\partial w_j} log(1-\sigma(x_i,w))$$

Y haciendo cuentas:

$$\frac{\partial l(w|X,y)}{\partial w_j} = \sum_{i=1}^{n} y_i \frac{\partial}{\partial w_j} log\left(\frac{\sigma(x_i,w)}{1-\sigma(x_i,w)}\right) + \frac{\partial}{\partial w_j} log(1-\sigma(x_i,w))$$

$$\frac{\partial l(w|X,y)}{\partial w_j} = \sum_{i=1}^{n} y_i \frac{\partial}{\partial w_j}\left(w_0 + \sum_{j=1}^{n} w_{ij}x_{ij}\right) + \frac{\partial}{\partial w_j} - log\left(1+\exp^{\left(w_0+\sum_{j=1}^{n} w_{ij}x_{ij}\right)}\right)$$

Aplicando las derivadas:

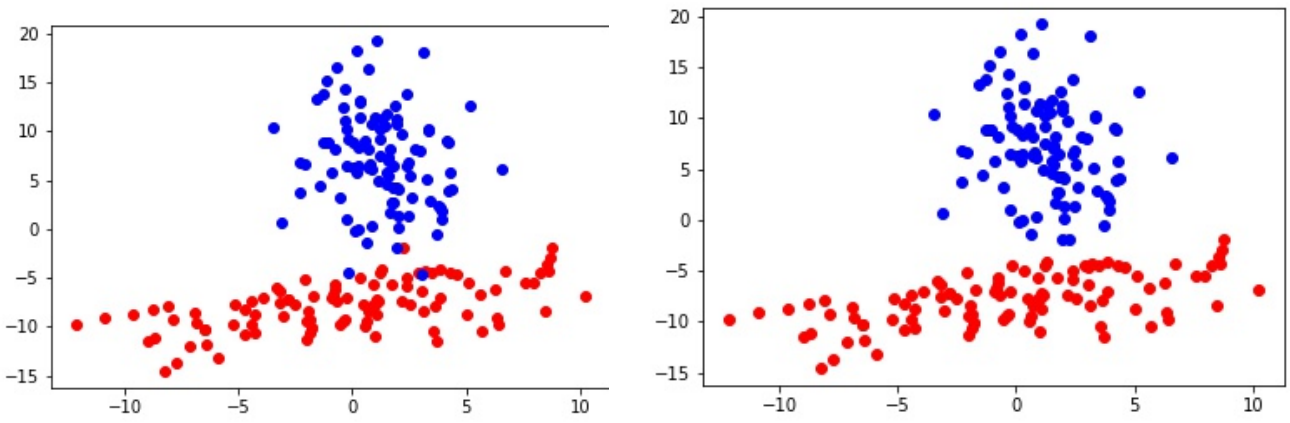
$$\frac{\partial l(w|X,y)}{\partial w_j} = \sum_{i=1}^{n} y_i x_{ij} - \frac{\exp^{\left(w_0+\sum_{j=1}^{n} w_{ij}x_{ij}\right)} x_{ij}}{1+\exp^{\left(w_0+\sum_{j=1}^{n} w_{ij}x_{ij}\right)}}$$

Y reduciendo entonces con la expresión que se conoces para $\sigma(x_i, w)$:

$$\frac{\partial l(w|X,y)}{\partial w_j} = \sum_{i=1}^{n} (y_i - \sigma(x_i,w))x_{ij}$$

5.3

Para los datos de prueba de la parte 2 se obtuvo una precision de 98,5 %.
A continuación se muestran los datos junto con los datos clasificados en la figura :



(a) Patrones no clasificados

(b) Patrones clasificados

Figura 6: Distribución de los patrones

Se observa que los datos que estaban superpuestos con la clase roja fueron clasificados como de la clase roja, mientras los que estaban mas cerca de los azules fueron clasificados como azules y fue en esos donde se introdujo el error.
También se procedió de la forma de set de entrenamiento y set de test. En todos los casos se ajustaron los parámetros para obtener el mejor valor en el set de prueba. Para este caso se obtuvo un valor también cercano a el obtenido anteriormente de un 98,3 % de precisión. En la figura 7 se muestran los resultados obtenidos donde se muestra que solo un valor de test fue clasificado erróneamente.

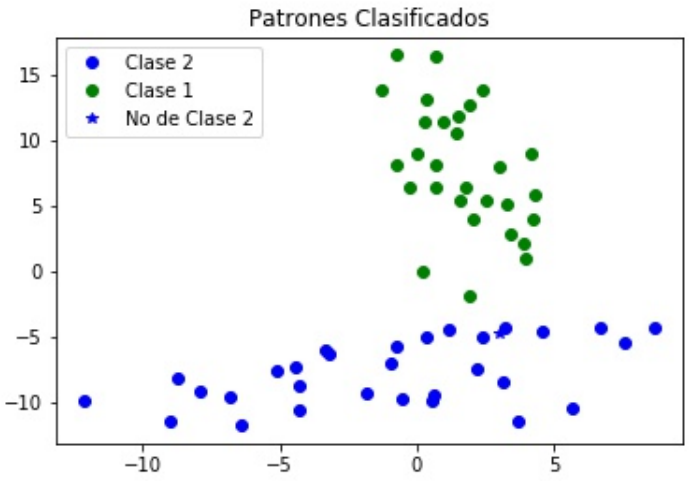


Figura 7: Datos separados por regresión logística en test set

Ejercicio 6

6.1

A continuación se muestran los patrones graficados para el ancho y el largo del pétalo en la figura 8.

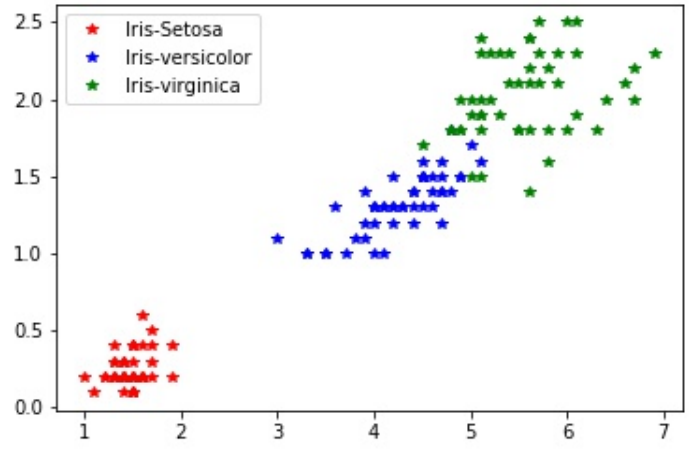
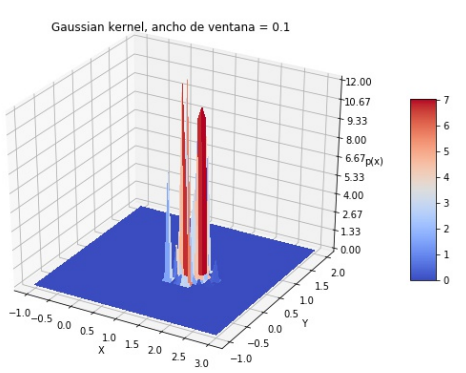


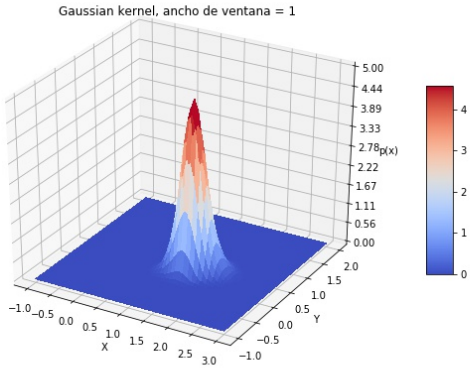
Figura 8

6.3

Se pudo apreciar que para valores muy bajos cerca de 0,1 las estimaciones se acercaban a deltas donde los datos están presentes, esto se debe a que el ancho de las gaussianas es reducido por lo cual no se puede hacer una buena estimación de $p(x)$. Lo mismo sucede para valores mucho mayores que 3 donde pasa lo contrario por lo cual la gaussiana es muy ancha dando una estimación mas expandida en el área. A continuación se muestran las gráficas para las distintas estimaciones:

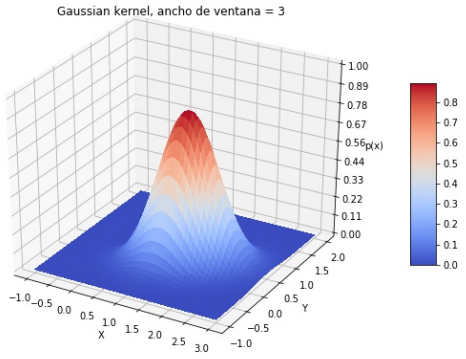


(a) Ancho de ventana 0,1

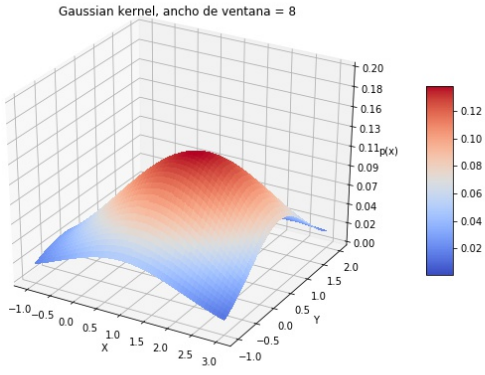


(b) Ancho de ventana 1

Figura 9: Distribución de la probabilidad para kernel gaussiano con distintos anchos de ventana



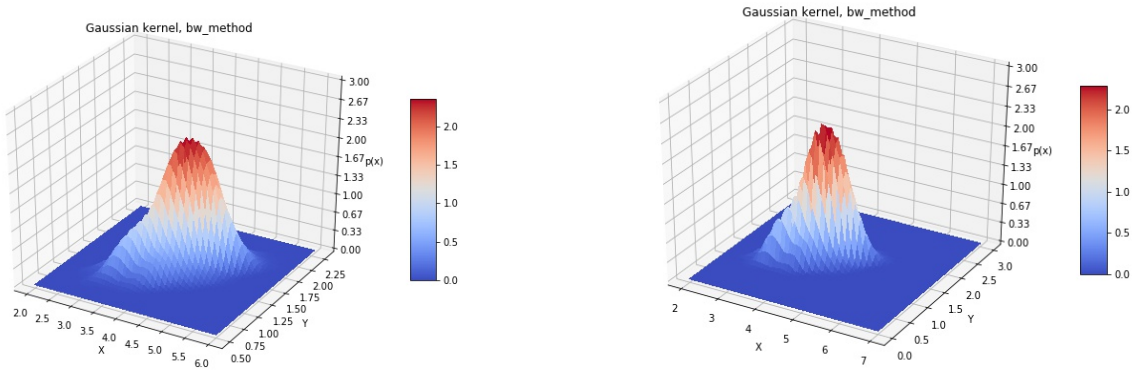
(a) Ancho de ventana 3



(b) Ancho de ventana 8

Figura 10: Distribución de la probabilidad para kernel gaussiano con distintos anchos de ventana

También se probaron con otros tipos de distancias predefinidas como lo son scott y silverman, con las que se pudo ver que la aproximación no fue tan suave como cuando se varia el ancho de la ventana de forma natural.



(a) Estimación con distancia de scott

(b) Estimación con distancia de silverman

Figura 11: Distribución de la probabilidad para kernel gaussiano con distintos anchos de ventana