Una introducción a los métodos númericos en el cálculo de estimadores máximo verosimiles

Andrés Mejía

Mayo 16, 2017

Introducción

Los sistemas de computo son una herramienta indispensable en el análisis moderno de datos, no solo nos permiten la manipulación de un gran volumen de información que de otra forma seria imposible de manejar, sino que tambien permite realizar cálculos que de otra manera serian imposible de realizar.

A pesar de esto no podemos ver el computador como una simple caja negra a la cual se le introducen datos y de la cual se obtiene una salida. El simple hecho de manejar un volumen moderado a grande de datos en un computador usual de la actualidad requiere que usemos de forma eficiente el poder computacional que tenemos. Si usamos R tal y como viene de forma estandar en la mayoria de distribuciones para PC, solo usaríamos uno de los varios procesadores que con seguridad contamos en computador moderno. Estamos limitados a que nuestros datos esten guardados en la memoria RAM (que rapidamente desborda muchas de las bases de datos en las que seria interesante trabajar).

La solución a este problema no es un computador mas grande y con mas RAM, para iniciar el poder del procesador esta allí inutilizada y en segundo lugar con la velocidad que se generan datos en la actualidad cualquier cantidad de memoria RAM se desbordará rapidamente.

Por esto cualquier sistema de "Big Data" debe ser capaz de repartir el trabajo entre varias entidades, llamense nucleos, computadores o tarjetas gráficas. El primer paso de esta distribución de trabajo es entender lo que se esta haciendo.

Problema Inicial

Supongamos que tenemos una población en que viene de dos subpoblaciones normales independientes, esta variable se distribuye en cada subpoblación con media y varianza diferentes. El resultado de esto es la mixtura.

El objetivo es estimar la media y la varianza de las subpoblaciónes, asi como la probabilidad de que un individuo pertenezca a alguna de las mismas.

‘ ## Problema módelo

Se mide una variable en una población que viene de dos subpoblaciones normales independientes, esta variable se distribuye en cada subpoblación con media y varianza diferentes. El resultado de esto es una mixtura dada por:

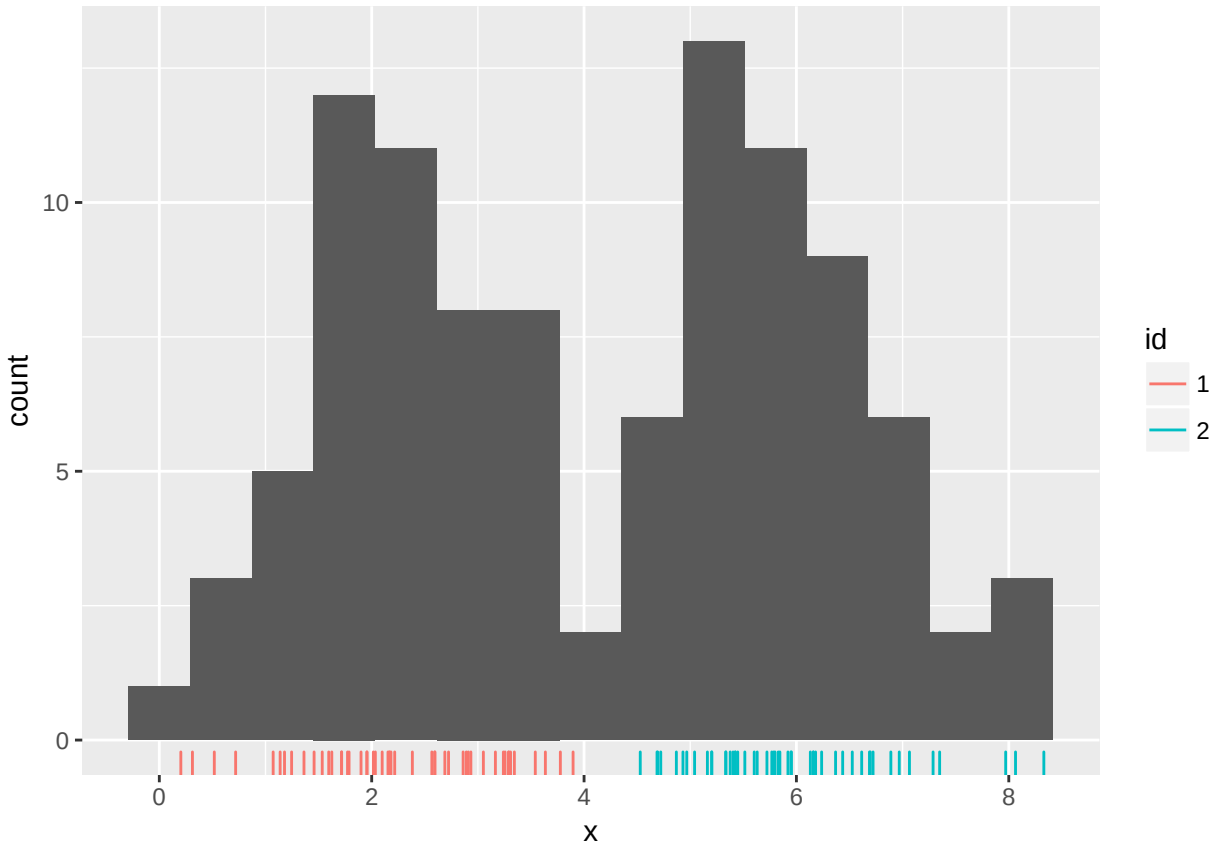
$$f(x;\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\pi) = \pi d(x;\mu_1,\sigma_1^2) + (1-\pi)d(x;\mu_2,\sigma_2^2)$$

Donde $\mu_1,\mu_2,\sigma_1^2,\sigma_2^2$ son mas medias y las varianzas de cada una de las subpoblaciones y π es la probabilidad de pertenecer a la subpoblación 1.

El objetivo es estimar la media y la varianza de las subpoblaciónes, así como el parámetro π

Es importante notar que no se tiene información de a que subpoblación pertenece cada individuo.

Se muestra a continuación unos datos simulados que corresponden a la situacion mencionada anteriormente con $\pi = 0.5$ $\mu_1 = 1$, $\mu_2 = 6$ y $\sigma_1^2 = \sigma_2^2 = 1$



Estimadores de Maximoverosimilitud.

La densidad del módelo es:

$$f(x;\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\pi) = \pi d(x;\mu_1,\sigma_1^2) + (1-\pi)d(x;\mu_2,\sigma_2^2)$$

Con esto la función de log-verosimilitud queda:

$$l(x;\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\pi) =$$

$$\sum_{i=1}^{n} \ln(\pi d(x_i;\mu_1,\sigma_1^2) + (1-\pi)d(x_i;\mu_2,\sigma_2^2))$$

Obteniendo un sistema a partir de las derivadas tenemos:

$$\frac{dl}{d\pi} = \sum_{i=1}^{n} \frac{d(x_i;\mu_1,\sigma_1^2) - d(x_i;\mu_2,\sigma_2^2)}{\pi d(x;\mu_1,\sigma_1^2) + (1-\pi)d(x;\mu_2,\sigma_2^2)}$$

$$\frac{dl}{d\mu_k} = \sum_{i=1}^{n} \frac{(-1)^{k+1}((3-2k)\pi + k - 1)\frac{d}{d\mu_k}d(x_i;\mu_k,\sigma_k^2)}{\pi d(x;\mu_1,\sigma_1^2) + (1-\pi)d(x;\mu_2,\sigma_2^2)}$$

$$\frac{dl}{d\sigma_k^2} = \sum_{i=1}^{n} \frac{(-1)^{k+1}((3-2k)\pi + k - 1)\frac{d}{d\sigma_k^2}d(x_i;\mu_k,\sigma_k^2)}{\pi d(x;\mu_1,\sigma_1^2) + (1-\pi)d(x;\mu_2,\sigma_2^2)}$$

Encontrar una solución análitica a este sistema de ecuaciones no parece factible por lo que tendremos que recurrir a métodos númericos.

El método de Newton

Sea $f(x)$ una función de la cúal deseamos encontrar una raiz, sea α esta raiz. Si realizamos la expansión de Taylor en el punto x_i

$$f(x) = f(x_i) + f'(x_i)(x - x_i) + O((x - x_i)^2)$$

Si evaluamos en α e ignoramos el segundo termino obtenemos:

$$0 = f(\alpha) = f(x_i) + f'(x_i)(\alpha - x_i)$$

Así

$$\alpha = x_i - \frac{f(x_i)}{f'(x_i)}$$

Este α que estamos proponiendo no es exactamente la raiz (ya que aproximamos al descartar los terminos de orden 2), pero esperamos que este más cerca de la raiz. El método entonces esta caracterizado por la sucesión:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Esto se generaliza a funciones $f : \mathbb{R}^{\aleph} \to \mathbb{R}^{\beth}$ (Método de Newton-Rhapson) con

$$\vec{x_{i+1}} = \vec{x_i} - J(\vec{x_i})^{-1} f(\vec{x_i})$$

Donde $J(\vec{x_i})^{-1}$ es la inversa de la matriz Jacobiana de la función f (que en nuestro ejemplo va de $f : \mathbb{R}^{\not\exists} \to \mathbb{R}^{\not\exists}$).

Este método tiene las siguientes ventajas

- Es aplicable para cualquier función que cumpla unas condiciones mínimas. (Difenciable en una vencidad de la raiz y con segunda derivada acotada)
- Converge de manera cuadratica cuando se cumplen supuestos de suavidad de la función (Aproximadamente duplica los digitos exactos en cada iteración)

Tambien tiene algunas dificultados que no lo hacen aplicable a la todalidad de situaciones

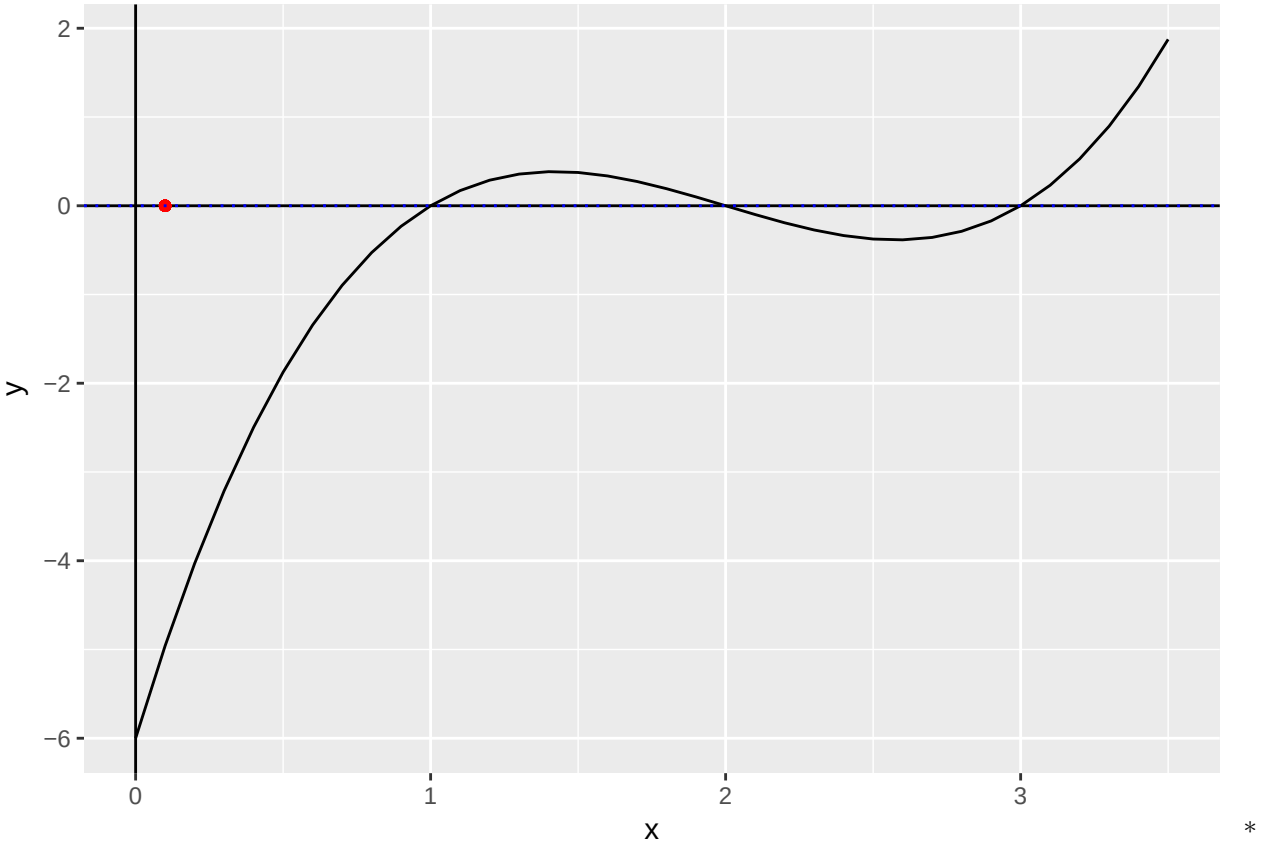
- Puede no converger si no se inicia con un punto adecuado (converge en una vecindad de la solución).
- Hay dificultad operacional al calcular la derivada (la matriz J y su inversa)
- Encuentra solo una solución (en caso de existir varias)
- No aprovecha la estructura del problema (optimización)

Las dificultades operacionales del método no son puntos menores, la inversión de una matriz es un problema especialmente complejo que puede hacer que perdamos lo ganado al tener convergencia cuadrática. En la practica no se encuentra la inversa de la matriz J, sino que se resulve un sistema de equaciones equivalente.

La complicación del cálculo de la derivada es relativa, dado que podemos aproximarla usando una ecuación de diferencia finita, esto tiene el efecto de reducir la convergencia y aumentar el número de operaciones. En la practica se espera que en la medida de lo posible se de la derivada de forma análitica al método para mejorar su velocidad de convergencia.

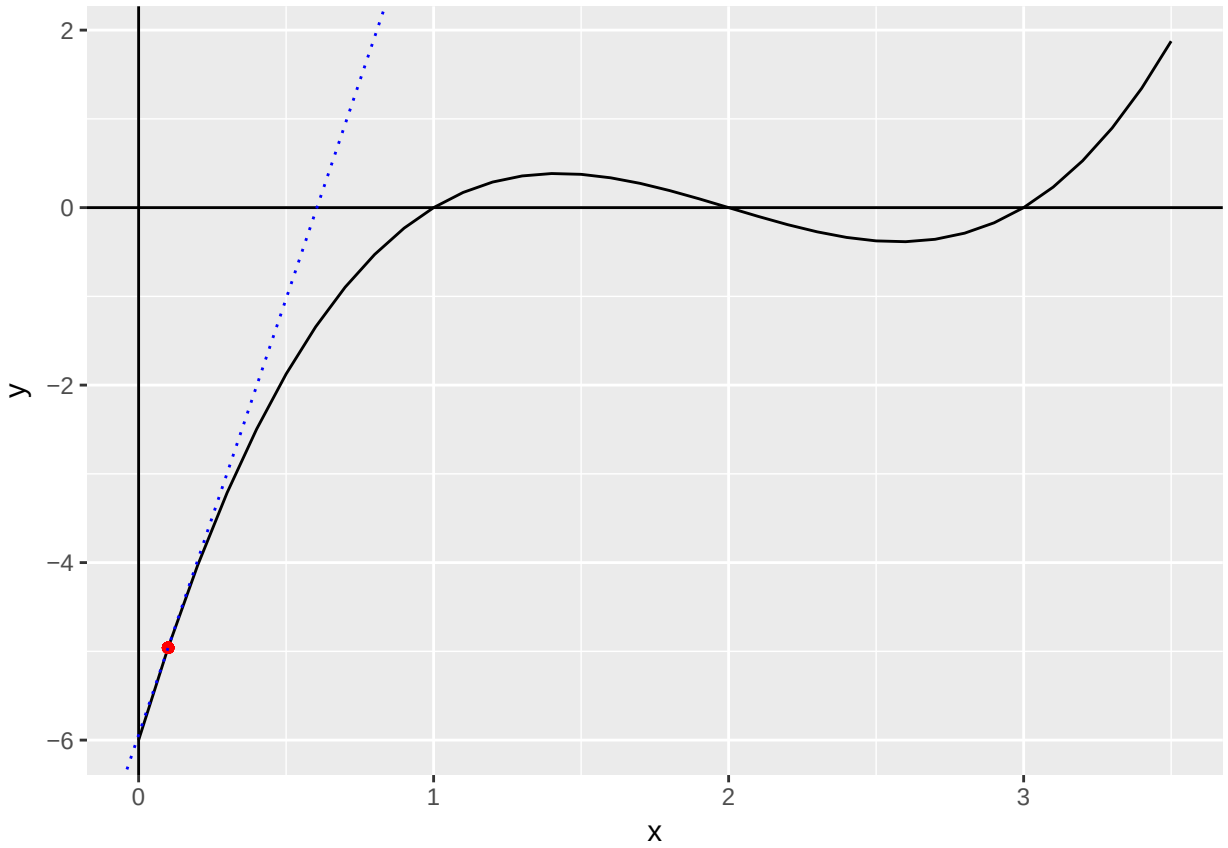
Representación gráfica del método

- Paso 0: Dar un valor inicial a x.

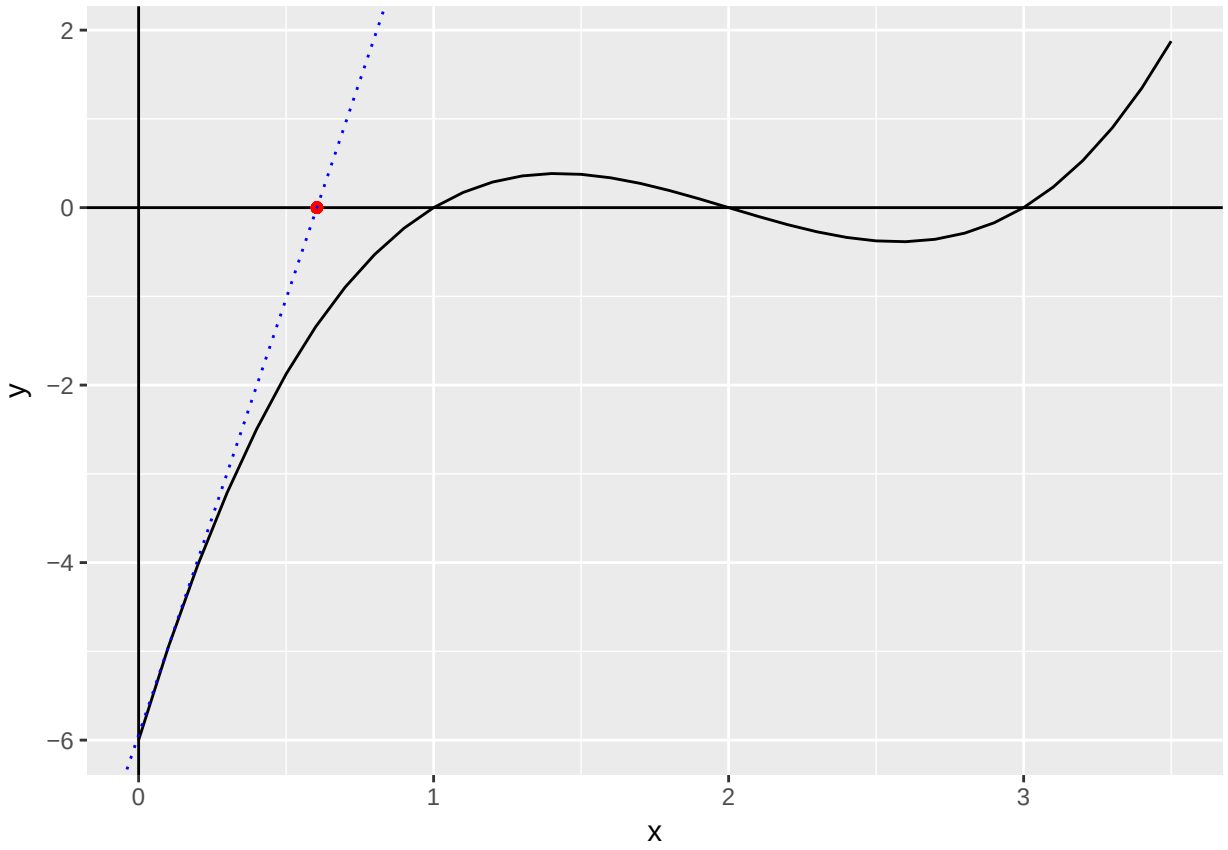


*

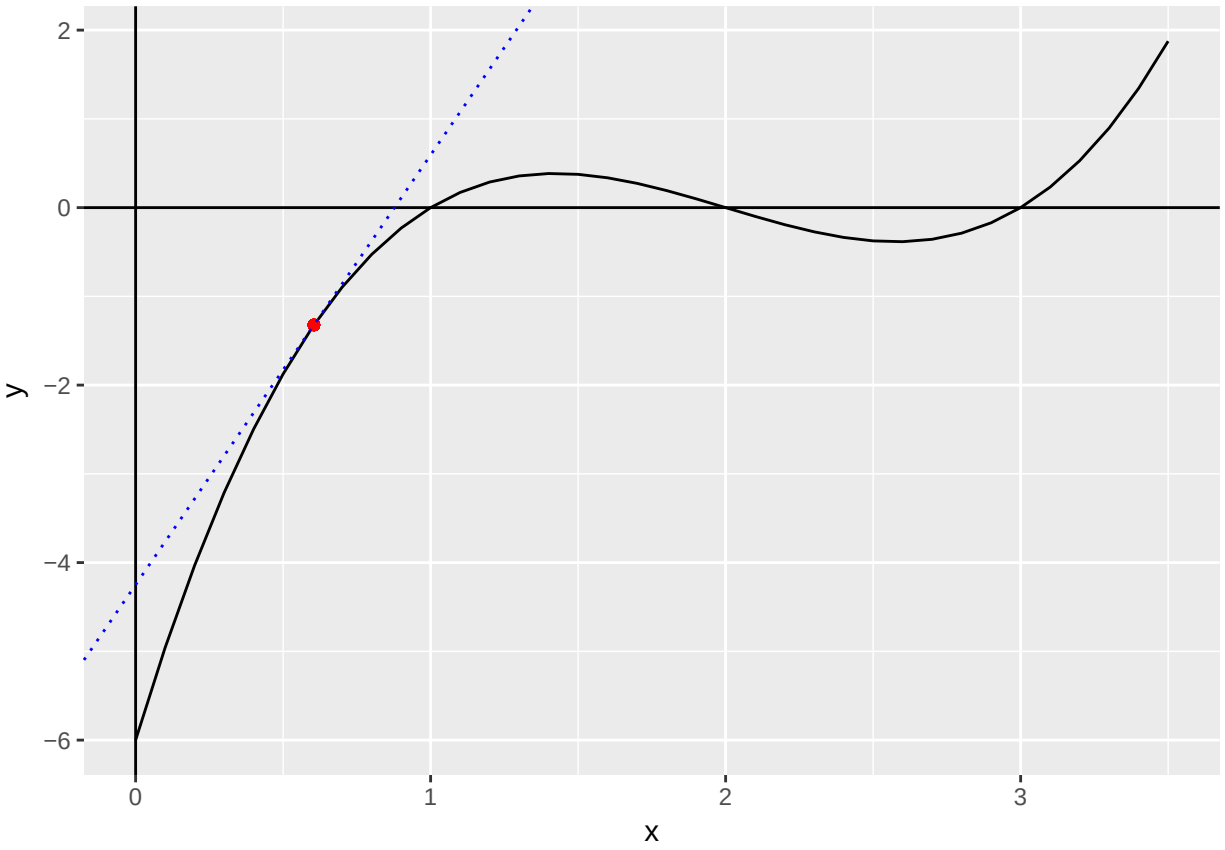
Paso 1: Calcular la recta tangente de la función en el punto dado.



- Paso 3: Usar la recta tangente para actualizar el x.



- Paso 4: Repetir hasta la convergencia.



Divergencia en el método de Newton

Como se meciono anteriormente una de las fallas del método es que para lograr la convergencia tenemos que estar ya relativamente cerca de la solución. Afecta mucho que alguna de las iteraciones este cerca de puntos con pendiente cero (o puntos donde la matriz J no sea invertible).

Tambien hay que considerar que especialmente cerca de estos puntos se puede dar un comportamiento caotico de las soluciones, es decir que con variaciones pequeñas del valor inicial se puede llegar a soluciones considerablemente distintas, el comportamiento al tratar de solucionar algunos sistemas incluso da lugar a cierto tipo de fractales llamados fractales de Newton.

Ejemplo: Programación en R

```
func=function(x){(x^3-6*x^2+11*x-6)*1}
dfunc=function(x){(3*x^2-12*x+11)*1}
valor=list(-8)
for(i in 2:10){
        valor[[i]]=valor[[i-1]]-func(valor[[i-1]])/dfunc(valor[[i-1]])
}
data.frame(unlist(valor))
```

```
##    unlist.valor.
## 1     -8.0000000
## 2     -4.6889632
## 3     -2.4927804
```

```
## 4     -1.0454795
## 5     -0.1060077
## 6      0.4819019
## 7      0.8168003
## 8      0.9646914
## 9      0.9982722
## 10     0.9999955
```

Método EM (Expectation Maximization)

Una de las criticas al uso del método de Newton para resolver este tipo de problemas es que no toma en cuenta su estructura como problema de optimización, asi como tampoco las propiedades estadisticas del problema. El método EM usa estas estructuras para llegar a un método alternativo con propiedades distintas al método de Newton.

Ejemplo

En el caso especial de una mixtura como en el ejemplo inicial.

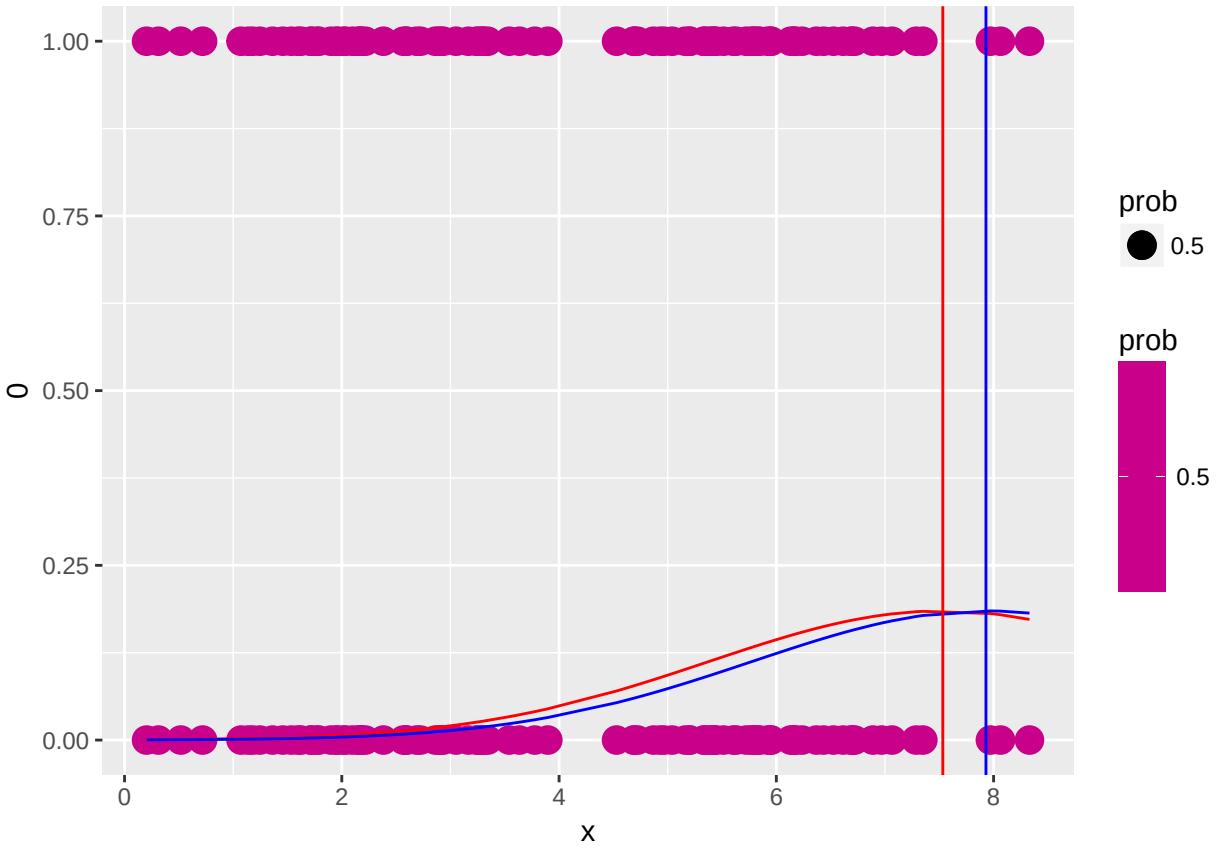
- Inicie los parametros $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2.\pi$
- Expectation Step: calcule las probabilidates

$$\gamma_i = \frac{\pi\Phi_{\mu_2}(y_i)}{\pi\Phi_{\mu_2}(y_i) + \pi\Phi_{\mu_1}(y_i)}$$

$$\pi = \bar{\gamma}_i$$

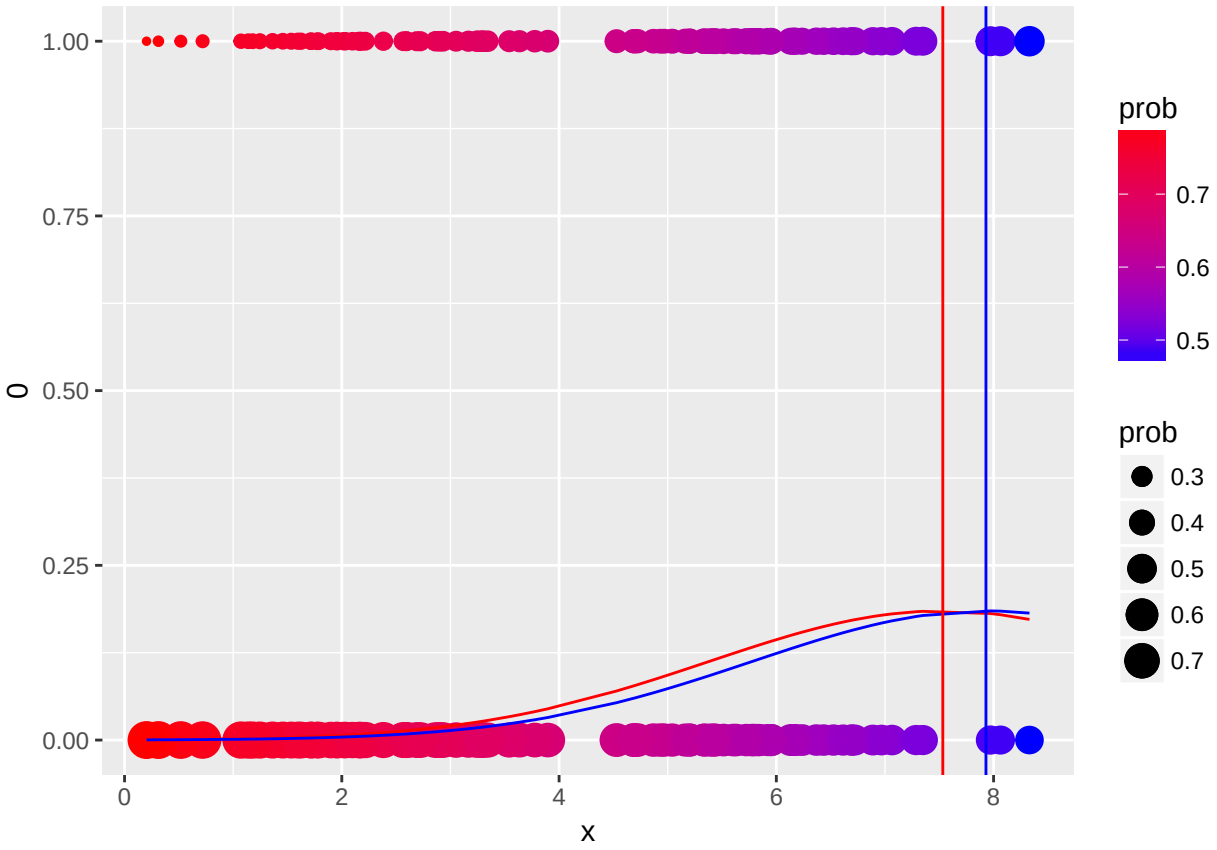
- Paso de Maximización: Obtenga el estimador maximo verosimil para $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$. Que son los estimadores usuales tomando como pesos las probabilidades γ_i calculadas en el paso anterior.
- Repita los pasos dos y tres hasta la convergencia

Para este ejemplo se utilizaron números aleatorios entre 0 y 10 para inicializar las medias, las varianzas se iniciaron cada una con la mitad de la varianza muestral.



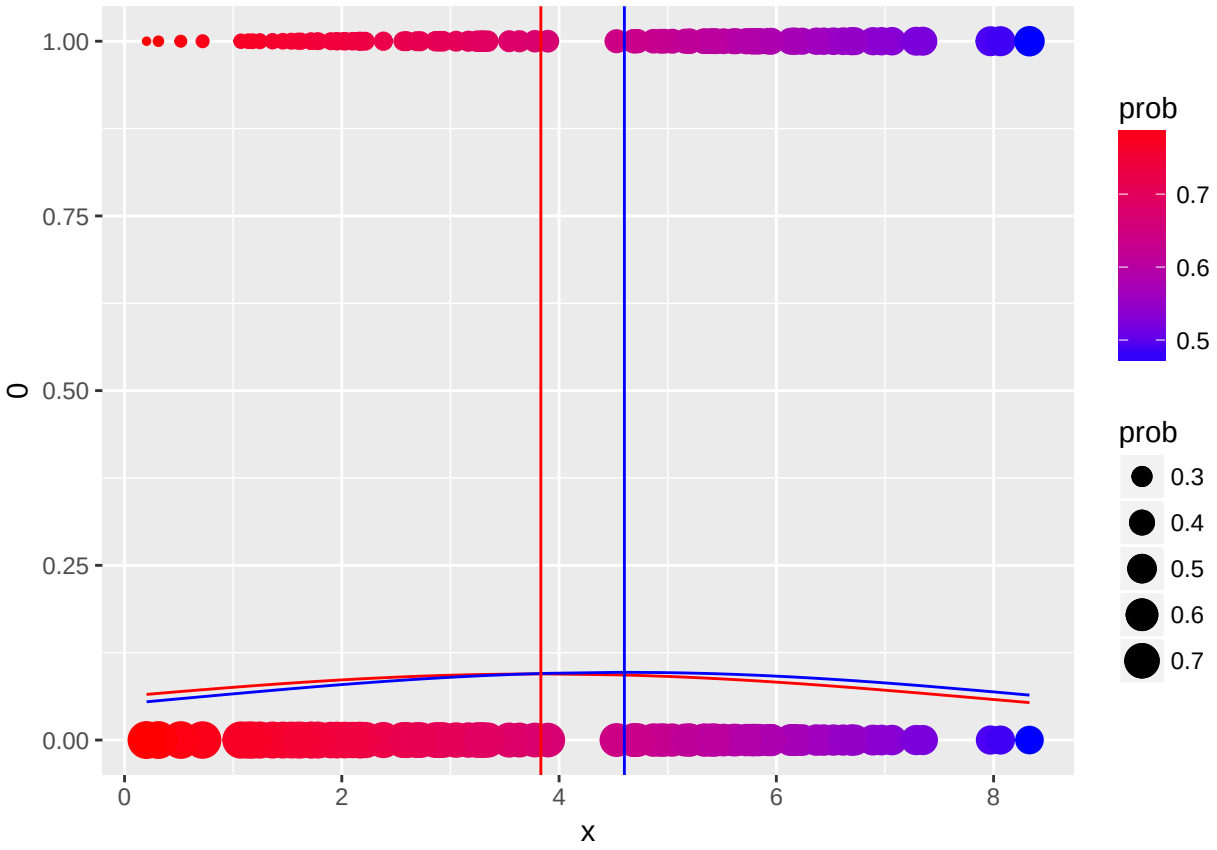
- Paso 1.

Se actualiza la probabilidad de que cada punto pertenezca a cada poblacion (es decir se calcula γ_i)



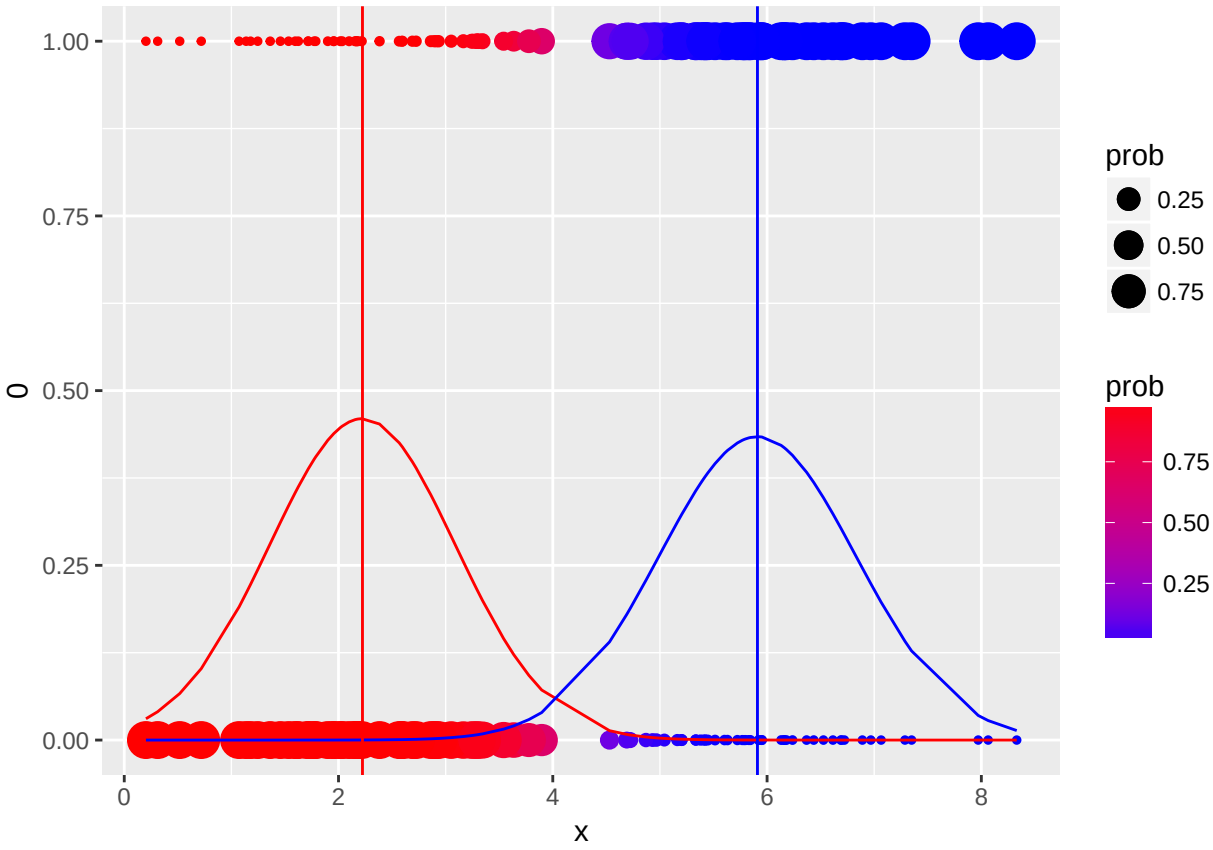
- Paso 2.

Con esta información se estiman de nuevo medias y varianzas usando el estimador maximo verosimil.



- Paso 3.

Repetir hasta la convergencia



Explicación

Consideremos primero una función de verosimilitud de la cual queremos estimar sus parametros θ y tenemos un conjunto de variables Z observadas, es decir $l(\theta, Z)$, consideremos ahora que existe un conjunto de variables que no observamos Z' (la pertenencia a una de las categorias en el ejemplo inicial), tambien sea θ' un estimador de θ. Y consideremos que expandimos nuestro problema inicial con estimaciones de estas variables, si:

$$P(Z'|Z,\theta) = \frac{P(Z', Z|theta')}{P(Z|\theta')}$$

de forma equivalente:

$$P(Z|\theta') = \frac{P(Z', Z|\theta')}{P(Z'|Z,\theta)}$$

Visto como verosimilitud y tomando esperanza con respecto a $Z', Z|Z$

$$l(\theta', Z) = E[l_0(\theta', Z, Z')|Z,)\theta] - E[l_1(\theta', Z|Z')|Z,)\theta]$$

$$l(\theta', P(Z')) = E_P[l_0(\theta', Z, Z')] - E_P[\log(P(Z'))]$$

Ahora si tomamos el maximizador sobre $P(Z')$ tomando θ fijo se puede encontrar en

$$P(Z') = Pr(Z'|Z, \theta')$$

Que es equivalente al paso E, el paso M es tomar $P(Z')$ fijo y tomar la maximización con respecto a θ' que corresponde a los estimadores de máxima verosimilitus usuales.

Propiedades

Este método cuenta con algunas ventajas con respecto al método de Newton.

- Siempre converge.
- Relativamente intuitivo y facil de implementar.

Sin embargo tiene su propio conjunto de incovenientes:

- Solo funciona para problemas de optimizacion que cumplen cierta estructura.
- La velocidad de convergencia es baja.
- Esta convergencia no se puede asegurar que sea a un optimo global.
- Encuentra solo una solución (en caso de existir varias).

A continuación.

Este análisis tambien abre la puerta a otra cantidad de métodos que aprovechen la extructura de optimización del problema (por ejemplo metodos de gradiente conjugado u otros mas heuristicos). Hay que considerar que