Tarea 8 Reconocimiento de Patrones

Yamil Ernesto Morfa Avalos

15 de mayo de 2021

B. Preguntas cortas

1. Supongamos que se usa la codificación $y \in \{-1,\, 1\}$

$$logit \frac{P\left(Y=1 \,|\, X=x\right)}{P\left(Y=0 \,|\, X=x\right)} = \sum_{t=1}^{n} \beta_t x_t + \alpha = x^T\beta + \alpha$$

$$P\left(y=1 \,|\, X=x\right) = \frac{\exp\left(x^T\beta+\alpha\right)}{1+\exp\left(x^T\beta+\alpha\right)}$$

y luego $P\left(y=-1 \,|\, X=x\right) = 1 - P\left(y=1 \,|\, X=x\right)$

$$P\left(y=-1 \,|\, X=x\right) = 1 - \frac{\exp\left(x^T\beta+\alpha\right)}{1+\exp\left(x^T\beta+\alpha\right)} = \frac{1}{1+\exp\left(x^T\beta+\alpha\right)} = \frac{1}{1+\exp\left(-y\left\{x^T\beta+\alpha\right\}\right)}$$

ahora veamos si se cumple que para $y = 1$

$$\frac{1}{1+\exp\left(-y\left\{x^T\beta+\alpha\right\}\right)} = \frac{1}{1+\exp\left(-\left\{x^T\beta+\alpha\right\}\right)} = \frac{1}{1+\exp\left(-\left\{x^T\beta+\alpha\right\}\right)} \frac{\exp\left(x^T\beta+\alpha\right)}{\exp\left(x^T\beta+\alpha\right)} =$$

$$= \frac{\exp\left(x^T\beta+\alpha\right)}{\exp\left(x^T\beta+\alpha\right)+1} = P\left(y=1 \,|\, X=x\right)$$

Entonces:

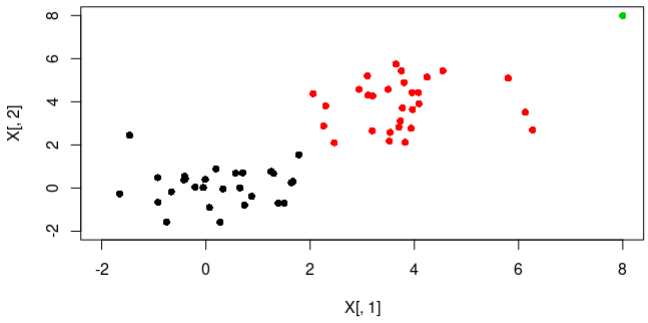
$$P\left(y \,|\, X=x\right) = \frac{1}{1+\exp\left(-y\left\{x^T\beta+\alpha\right\}\right)}$$

Luego la verosimilitud correspondiente es:

$$l\left(\beta,\, \alpha\right) = \log\left\{\prod_{t=1}^{n} P\left(y \,|\, X=x_t\right)\right\} = \sum_{t=1}^{n} \log \frac{1}{1+\exp\left(-y\left\{x_t^T\beta+\alpha\right\}\right)}$$

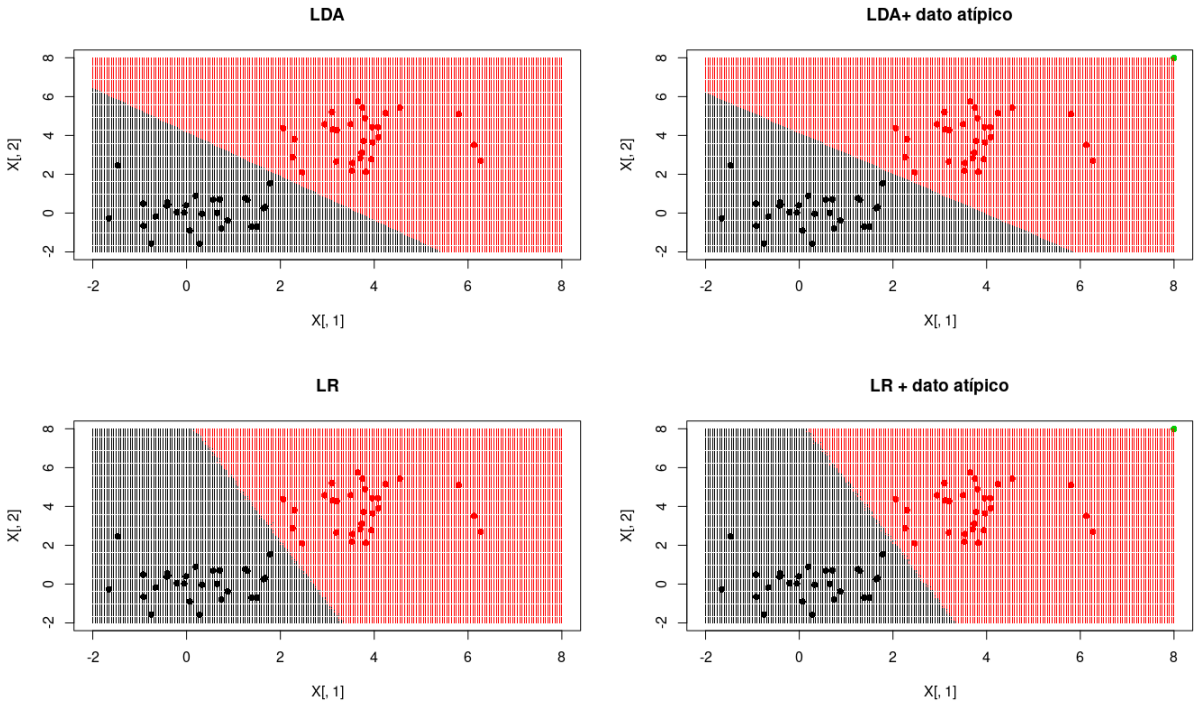
2. Consideremos las siguientes muestras de dos distribuciones normales con misma matriz de covarianza pero medias diferente

Figura 1:



El punto representado en rojo es una observación atípica de la clase azul.

Figura 2:



En la figura 2 se muestran las fronteras de decisión para estos datos con LDA y LR respectivamente, aunque en ninguno de los dos casos se aprecia una diferencia muy representativa, si podemos observar que la frontera de LDA varía. Según esto podemos suponer de la Regresión logística es más robusta a datos atípicos.

Esto tiene sentido, dado que LDA tiene como frontera de decisión $w^T x - w_0 = 0$ donde $w = \Sigma^{-1}(\mu_2 - \mu_1)$ y

$w_0 = \log \frac{\pi_1}{\pi_2} + \frac{1}{2}(\mu_2 + \mu_1)^T \Sigma^{-1}(\mu_2 - \mu_1)$ como podemos ver los valores estimados de las medias y la matriz de covarianza muestra influyen de manera directa en los parámetros de la frontera de decisión y estos valores son sensibles a datos atípicos, especialmente la media. En el ejemplo que se muestra los valores de las medias estimados fueron $\mu_1 = [0{,}26, -0{,}09]$ y $\mu_2 = [3{,}74, 3{,}88]$ para los datos normales y $\mu_1 = [0{,}26, -0{,}09]$ y $\mu_2 = [3{,}88, 4{,}01]$ para las normales mas el dato atípico. Como podemos ver μ_2 varía considerablemente.

Sin embargo en RL la clasificación viene dada por una binomial con parámetros $p_i = \frac{1}{1+\exp\left(-\left\{x_i^t\beta\right\}\right)}$ (incluyendo en cada x_i una componente constante 1) y n_i (numero de ensayos bernoulli, que se relaciona con la cantidad de clases. En nuestro caso $n_i = 1$), donde β se obtienen de maximizar la logverosimilitud $l(\beta) = \sum_i (y_i \beta^t x_i - \log(1 + \exp(\beta^t x_i)))$ dado que se consideran todos los datos, un dato atípico no influye demasiado en la estimación de β. En el ejemplo anterior se estimó $\beta = [-128{,}87,\ 47{,}72,\ 15{,}14]$ para los datos normales y $\beta = [-127{,}87,\ 47{,}27,\ 15{,}11]$.

C. Análisis de datos

1. Sea $Y \sim Bern(0{,}5)$, $X \mid Y = 0 \sim N([0,0], A)$, $X \mid Y = 1 \sim N([1,2], A)$ con:

$$A = \begin{bmatrix} 1 & 0{,}4 \\ 0{,}4 & 1 \end{bmatrix}$$

a. Sea $L(y, \hat{y}) = L(\hat{y}, y)$ una función de costo simétrica, tenemos que:

$$E_{Y|X=x}\left[L\left(y, \hat{y}(x)\right)\right] = L\left(0, \hat{y}(x)\right) P(Y = 0|X = x) + L\left(1, \hat{y}(x)\right) P(Y = 1|X = x)$$

$$E_{Y|X=x}\left[L\left(y, \hat{y}(x)\right)\right] = L\left(0, \hat{y}(x)\right) \frac{P(X = x|Y = 0) P(Y = 0)}{P(X = x)} + L\left(1, \hat{y}(x)\right) \frac{P(X = x|Y = 1) P(Y = 1)}{P(X = x)} =$$

$$= \frac{1}{P(X = x)} \left\{\pi_0 L\left(0, \hat{y}(x)\right) P(X = x|Y = 0) + \pi_1 L\left(1, \hat{y}(x)\right) P(X = x|Y = 1)\right\} =$$

$$= \frac{1}{P(X = x)} \left\{\pi_0 L\left(0, \hat{y}(x)\right) N([0,0], A) + \pi_1 L\left(1, \hat{y}(x)\right) N([1,2], A)\right\}$$

Si $\hat{y}(x) = 0$ entonces: $E_{Y|X=x}\left[L\left(y, \hat{y}(x)\right)\right] = \frac{1}{P(X=x)} \left\{\pi_1 L_{1,0} N([1,2], A)\right\}$

Si $\hat{y}(x) = 1$ entonces: $E_{Y|X=x}\left[L\left(y, \hat{y}(x)\right)\right] = \frac{1}{P(X=x)} \left\{\pi_0 L_{0,1} N([0,0], A)\right\}$

Luego:

$$\frac{1}{P(X = x)} \left\{\pi_1 L_{1,0} N([1,2], A)\right\} < \frac{1}{P(X = x)} \left\{\pi_0 L_{0,1} N([0,0], A)\right\}$$

$$\frac{\pi_1}{2\pi\sqrt{\det(A)}} \exp\left\{-\frac{1}{2}(x - \mu_1)^T A^{-1}(x - \mu_1)\right\} < \frac{\pi_0}{2\pi\sqrt{\det(A)}} \exp\left\{-\frac{1}{2}(x - \mu_0)^T A^{-1}(x - \mu_0)\right\}$$

$$\pi_1 \exp\left\{-\frac{1}{2}\left(x-\mu_1\right)^T A^{-1}\left(x-\mu_1\right)\right\} < \pi_0 \exp\left\{-\frac{1}{2}\left(x-\mu_0\right)^T A^{-1}\left(x-\mu_0\right)\right\}$$

dado que conocemos que $\mu_0 = [0,0]$

$$\log \pi_1 - \frac{1}{2}\left(x-\mu_1\right)^T A^{-1}\left(x-\mu_1\right) < \log \pi_0 - \frac{1}{2}x^T A^{-1} x$$

$$-\frac{1}{2}x^T A^{-1}x + \frac{1}{2}\mu_1^T A^{-1}x + \frac{1}{2}x^T A^{-1}\mu_1 - \frac{1}{2}\mu_1^T A^{-1}\mu_1 < \frac{\log \pi_0}{\log \pi_1} - \frac{1}{2}x^T A^{-1}x$$

$$\frac{1}{2}\mu_1^T A^{-1}x + \frac{1}{2}x^T A^{-1}\mu_1 - \frac{1}{2}\mu_1^T A^{-1}\mu_1 < \frac{\log \pi_0}{\log \pi_1}$$

Por ser A simétrica:

$$\mu_1^T A^{-1}x < \frac{\log \pi_0}{\log \pi_1} + \frac{1}{2}\mu_1^T A^{-1}\mu_1$$

Luego nuestro clasificador sería:

$$y^*(x) = \begin{cases} 0 & \mu_1^T A^{-1}x < \frac{\log \pi_0}{\log \pi_1} + \frac{1}{2}\mu_1^T A^{-1}\mu_1 \\ 1 & else \end{cases}$$

Dado que generamos la misma cantidad de datos para ambas normales, entonces $\pi_0 = \pi_1 = \frac{1}{2}$

$$y^*(x) = \begin{cases} 0 & \mu_1^T A^{-1}x < \frac{1}{2}\mu_1^T A^{-1}\mu_1 \\ 1 & else \end{cases}$$

b. Una normal multivariada se describe como:

$$N(x, \mu, A) = \frac{1}{(2\pi)^{n/2}\det^{1/2}(A)} \exp\left\{-\frac{1}{2}\left(x-\mu\right)^T A^{-1}\left(x-\mu\right)\right\}$$

Para el caso de una bivariada con $\mu = [0,0]^T$ esto se puede reescribir de la forma:

$$N([x,y], A) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y}\right)\right\}$$

con

$$A = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

luego:

$$A = \begin{bmatrix} 1 & 0{,}4 \\ 0{,}4 & 1 \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \Longrightarrow \sigma_x^2 = \sigma_y^2 = 1;\ \rho = 0{,}4$$

Entonces podemos obtener la primera normal bivariada generando dos normales estándar cuya combinación es una bivariada, cuya correlación ρ aleatoria. Luego podemos realizar una rotación de uno de estos ejes de manera tal que

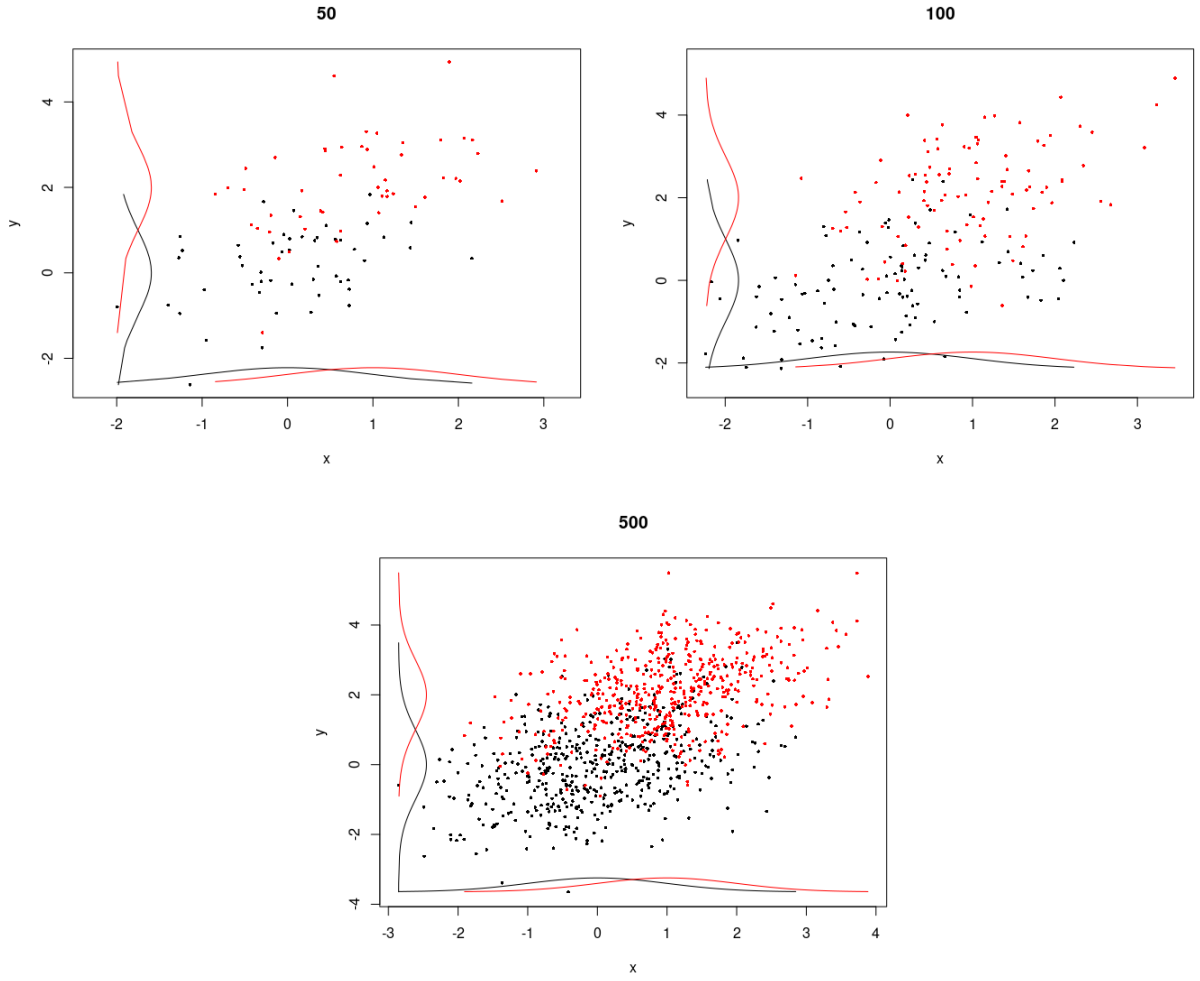
obtengamos la correlación deseada $\rho = 0{,}4$. Notemos que restringidos a la recta como subespacio 1 dimensional de $\mathbb{R}^2$ se mantiene la estructura normal de los datos.

Sea x_1 y x_2 dos vectores ortonormales, sea $x_3 = x_2 + \frac{x_1}{\tan\theta}$, entonces el angulo comprendido entre x_1 y x_3 es θ. Esto se puede demostrar fácilmente, $x_1^T x_3 = \|x_1\| \|x_3\| \cos(\alpha) \implies x_1^T x_2 + \frac{1}{\tan(\theta)} \|x_1\|^2 = \|x_1\| \|x_3\| \cos(\alpha) \implies \frac{1}{\tan(\theta)} = \frac{\|x_3\|}{\|x_1\|} \cos(\alpha) \implies \frac{1}{\tan(\theta)} = \frac{1}{\sin(\alpha)} \cos(\alpha) = \frac{1}{\tan(\alpha)} \implies \alpha = \theta$.

Para el caso $\mu_2 = [1, 2]$, se sigue el mismo proceso anterior para generar dos normales estándar con dicha correlación y luego basta con trasladarlas a las medias correspondientes.

Siguiendo este procedimiento se obtuvo las siguientes graficas que se muestran en la figura

Figura 3:

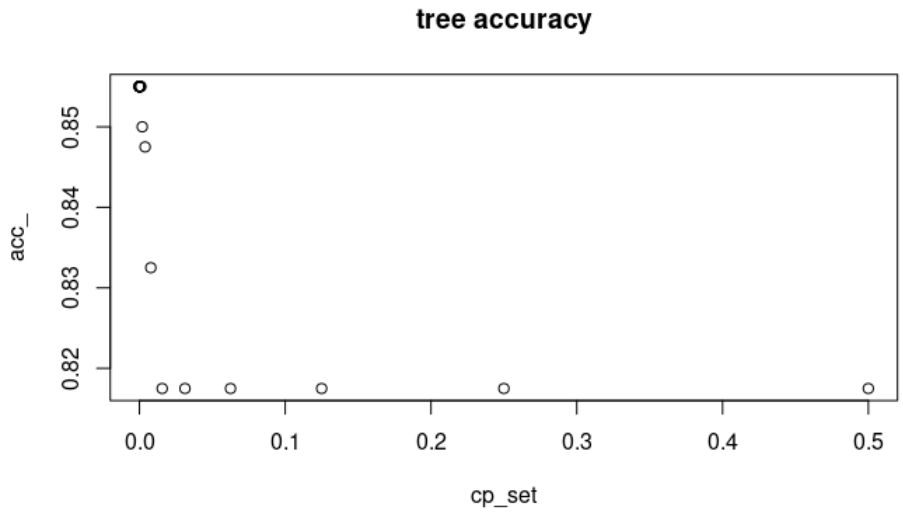


c. Para encontrar el mejor árbol generamos una muestra de 200 puntos de cada normal bivariada, teniendo un total de 400 observaciones. Usamos un árbol de clasificación binaria y variamos el parámetro que penaliza el tamaño del

árbol en la función de costo, *cp*. La métrica que represente la calidad del modelo que usaremos es:

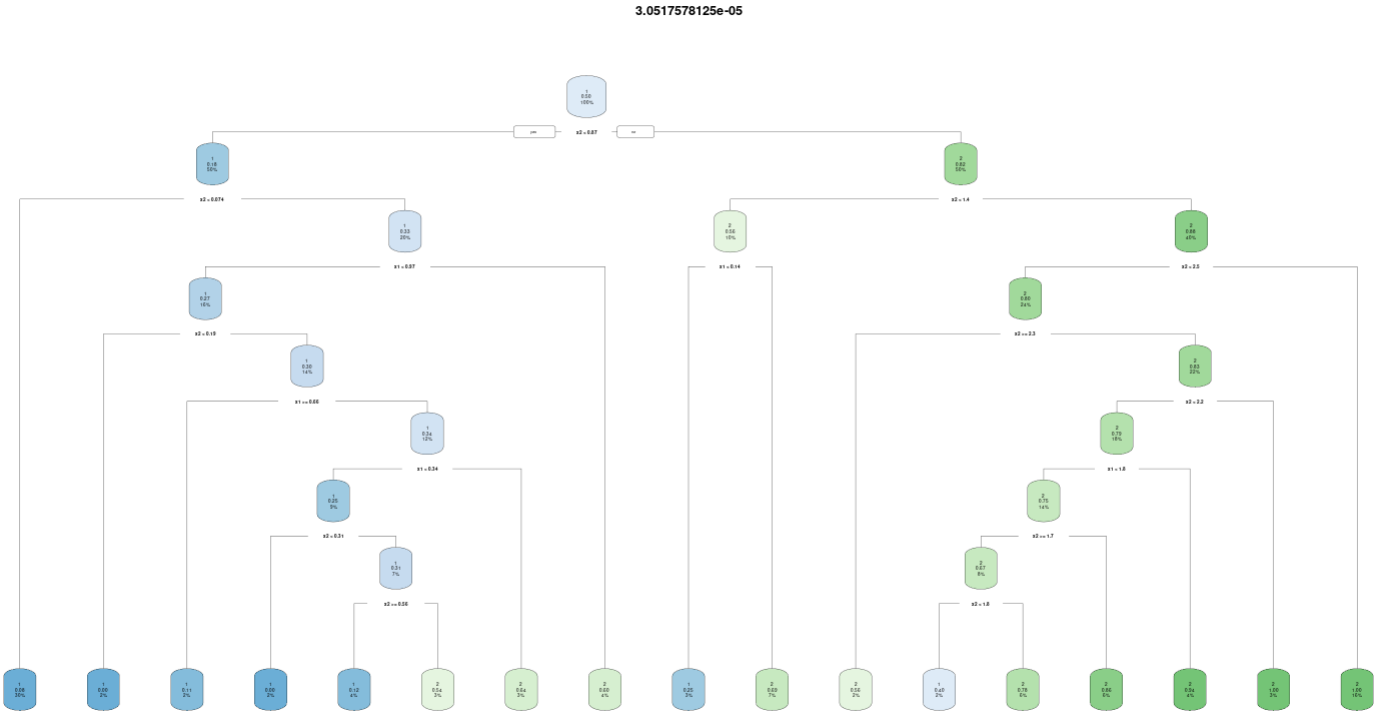
$$acc = \frac{\#\, predicciones\, correctas}{\#\, total\, predicciones}$$

Figura 4:



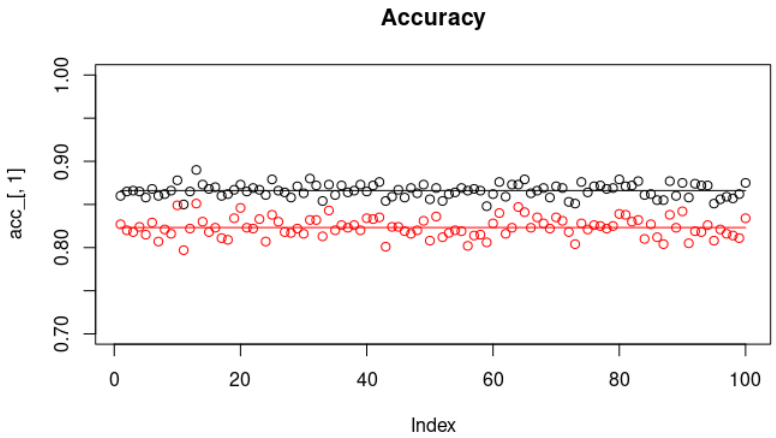
como podemos ver para valores mas bajos del parámetro *cp* se obtiene mejor accuracy, Además a pesar de penalizar tan fuertemente el tamaño del árbol (dejar que crezca), no se obtiene una estructura tan complicada:

Figura 5:



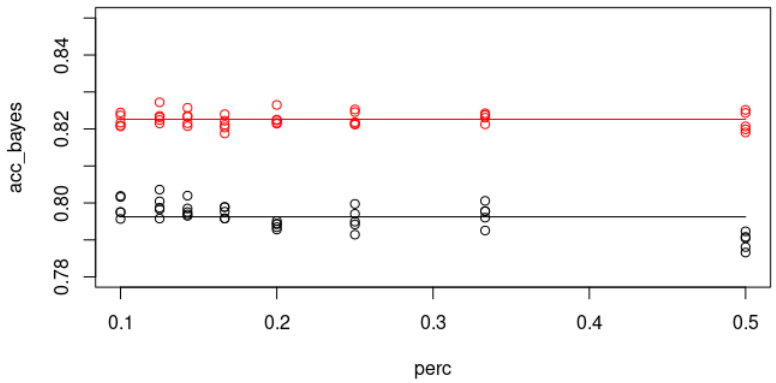
En la figura siguiente se muestra en rojo los valores de accuracy para el clasificador optimo de Bayes obtenido para distintas muestras generadas. En negro se muestra el árbol anterior.

Figura 6:



La validación cruzada es una técnica que nos permite evaluar la calidad de un modelo. Primero, dividimos el conjunto de datos de entrenamiento en B subconjuntos, llamados folds. Dato que para este ejemplo se realizará una validación cruzada completa. O sea se considerarán todos los datos del conjunto, el porcentaje de datos que se usará para la validación depende de B. Una vez dividido el conjunto se entrena el modelo con $train[i,]$ y se calcula la accuracy para los valores de $test\,[i,]$ con $i = 1, 2, ..., B$. Estos valores obtenidos de acc se promedian y esto es usado como métrica final para la calidad de el modelo.

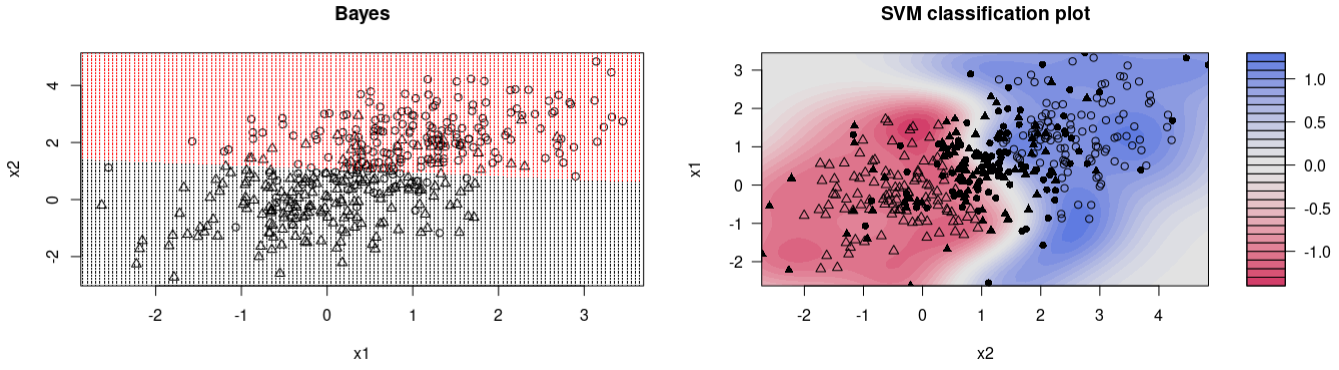
Figura 7:



El la figura 7 se muestra los resultados de la validación cruzada para los modelos de árbol con el parámetro cp obtenido y bayes en negro y rojo respectivamente.

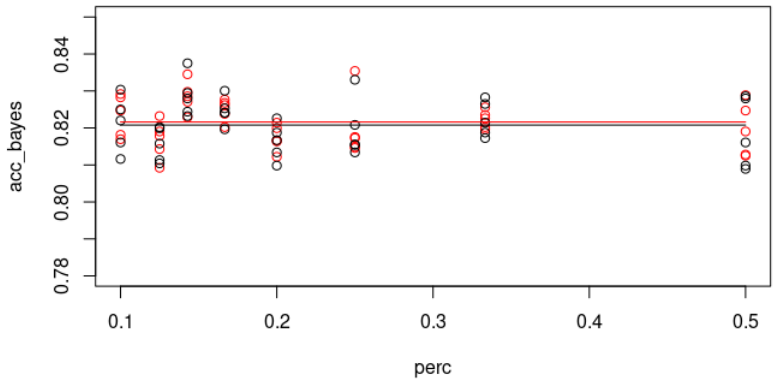
d. Se ha obtenido una Máquinas de soporte vectorial con kernel radial para estos datos y en la figura 8 se compara los modelos obtenidos con Bayes y KSVM de manera visual

Figura 8:



Además aplicamos validación cursada:

Figura 9:



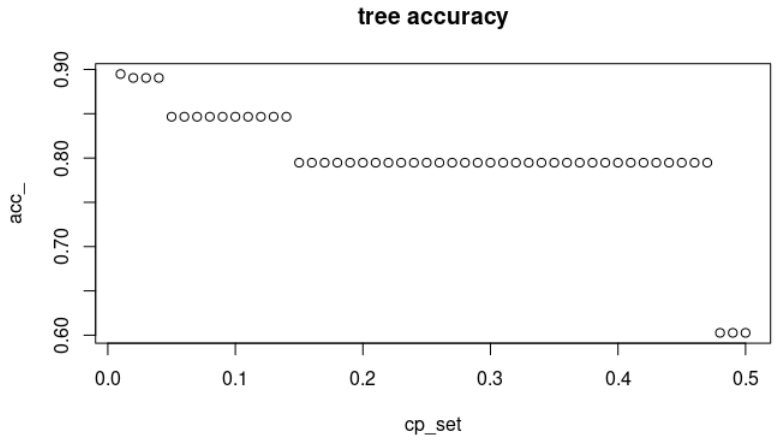
En la figura arriba se muestra los resultados de la validación cruzada, en rojo las accuracy de Bayes y en negro las de KSVM. Como se puede observar Bayes se comporta mejor para estos datos.

2. Trabajamos con los datos spam de http://search.r-project.org/library/kernlab/html/spam.html

El conjunto de datos contiene 2788 correos electrónicos clasificados como "no spam" y 1813 clasificados como "spam". Obteniendo un conjunto de datos con 4601 observaciones y 58 variables. Cada una de las primeras 57 variables nos da una información especifica con el fin de determinar si el correo es o no spam, cuya etiqueta aparece en la variable 58.

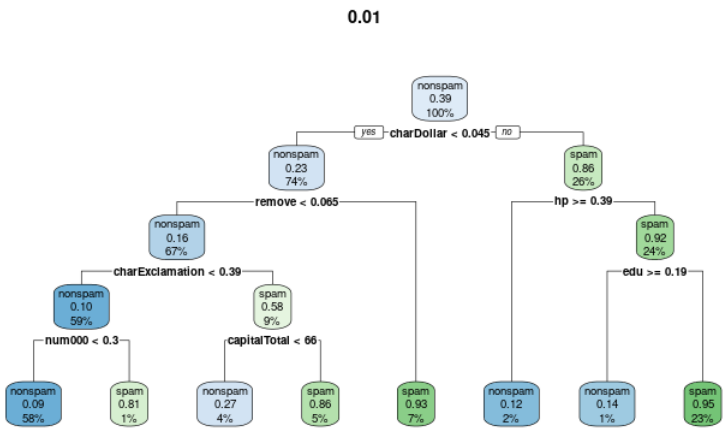
Primero busquemos un árbol de decisión adecuado. Como es un problema de clasificación binaria, usaremos los mismos arboles del ejercicio anterior y veremos con qué parámetro *cp* se obtiene mayor accuracy

Figura 10:



Para $cp = 0{,}01$ se obtuvo el mayor valor de $acc = 0{,}90$. El árbol resultante es:

Figura 11:



Podríamos seguir disminuyendo este valor y seguir encontrando arboles mas complejos, pero dado que se alcanza una buena precisión tomaremos este.

Para SVM con Kernel de base radial obtenemos una accuracy de 0,92 para estos mismos datos. Esta superioridad en precisión se confirma en la figura siguiente donde se generaron juegos de datos *train* y *test* distintos y se realizó el entrenamiento y clasificación con ambos modelos. En azul se representa KSVM y en rojo el árbol binario con parámetro $cp = 0{,}01$.

Figura 12: