SESSION 5 SUMMARY

Victor Miguel Terrón Macias

24/1/2021

SESION 5. REGRESION LINEAL Y CLASIFICACIÓN

INTRODUCCIÓN

Supongamos que nuestro trabajo consiste en aconsejar a un cliente sobre cómo mejorar las ventas de un producto particular, y el conjunto de datos con el que disponemos son datos de Publicidad que consisten en las ventas de aquel producto en 200 diferentes mercados, junto con presupuestos de publicidad para el producto en cada uno de aquellos mercados para tres medios de comunicación diferentes: TV, radio, y periódico. No es posible para nuestro cliente incrementar directamente las ventas del producto. Por otro lado, ellos pueden controlar el gasto en publicidad para cada uno de los tres medios de comunicación. Por lo tanto, si determinamos que hay una asociación entre publicidad y ventas, entonces podemos instruir a nuestro cliente para que ajuste los presupuestos de publicidad, y así indirectamente incrementar las ventas.

En otras palabras, nuestro objetivo es desarrollar un modelo preciso que pueda ser usado para predecir las ventas sobre la base de los tres presupuestos de medios de comunicación. En este contexto, los presupuestos de publicidad son las variables de entrada mientras que las ventas es una variable de salida. Las variables de entrada típicamente se denotan usando el símbolo X, con un subíndice para distinguirlas. Así X1 puede ser el presupuesto para TV, X2 el presupuesto para radio, y X3 el presupuesto para periódico. Las entradas tienen diferentes nombres, tales como predictores, variables independientes, características, o a veces solo variables. La variable de salida -en este caso, las ventas- frecuentemente es llamada la variable de respuesta o dependiente, y se denota típicamente con el símbolo Y.

Más generalmente, suponga que observamos una respuesta cuantitativa Y y p diferentes predictores, X1, X2, . . . , Xp. Asumimos que hay alguna relación entre Y y X=(X1, X2, . . . , Xp), la cual podemos escribir en la forma muy general

$$Y = f(X) + \varepsilon$$

Figure 1: FORMULA

Aquí f es alguna función desconocida pero fija de X1, X2, . . . , Xp, y es un término de error aleatorio, el cual es independiente de X y tiene media cero. En esta formulación, f representa la información sistemática que X proporciona acerca de Y. Sin embargo, la función f que conecta las variables de entrada a la variable de salida en general es desconocida. En esta situación debemos estimar f basados en los datos observados. En esencia, el aprendizaje estadístico se refiere a un conjunto de enfoques para estimar f.

¿POR QUÉ ESTIMAR f?

Hay dos razones principales por las cuales podemos desear estimar f: predicción e inferencia.

REGRESION LINEAL SIMPLE

Con frecuencia es necesario determinar si dos variables (aleatorias) están relacionadas de alguna manera. Por ejemplo, ¿tendrán los años de educación efecto sobre el salario que percibe un individuo? La relación entre dos variables cuantitativas puede visualizarse en un diagrama de dispersión en el plano, representando los valores de las variables en los ejes horizontal y vertical.

La correlación puede darse entre variables sin ninguna implicación de causalidad entre ellas, por ejemplo: si tomamos una muestra de individuos y medimos los diámetros del antebrazo y del muslo, seguramente encontraremos que hay una correlación positiva alta. Evidentemente no hay ninguna relación de causalidad entre estas variables y más bien ambas dependen del peso y la altura del individuo. A este tipo de correlación entre variables se le conoce como correlación espuria. La asociación más simple entre variables es cuando éstas se relacionan en forma lineal, sin embargo, no siempre es posible establecer este tipo de relación entre ellas. Para medir la magnitud de la asociación lineal entre dos variables, se utiliza comúnmente el coeficiente de correlación introducido por Karl Pearson. Éste es un número entre el -1 y el 1 denotado por la letra R. Si R = -1, se tiene una relación negativa perfecta y los puntos en el diagrama de dispersión se encuentran sobre una recta con pendiente negativa. Si R = 1, la relación lineal es también perfecta pero positiva: los puntos en el diagrama de dispersión están sobre una recta con pendiente positiva. Si R = 0, entonces no hay relación lineal alguna y los puntos forman más bien una nube difusa o algún otro patrón evidentemente no lineal. Lo usual es tener casos intermedios, en donde existe algún grado moderado de correlación lineal entre las variables. En general, en las ciencias sociales es raro tener coeficientes de correlación mayores que 0.7 (o menores que -0.7). A continuación, tenemos datos de estatura y pesos de unos individuos. Altura <- c(1.94, 1.82, 1.75, 1.80, 1.62, 1.64, 1.68, 1.46, 1.50, 1.55, 1.72, 1.67, 1.57, 1.60) Peso <- c(98, 80, 72, 83, 65, 70, 67, 47, 45, 50, 70, 61, 50, 52) Para obtener el coeficiente correlación de Pearson únicamente ejecutamos la siguiente instrucción en R cor(Altura, Peso) lo cual nos da 0.9645. A continuación, vamos a ajustar un modelo de regresión lineal simple a un conjunto de datos en R. Suponga que el conjunto de datos proviene de una fábrica que elabora productos

Para cada caso se considera un tamaño del proceso o tamaño de la ejecución (RunSize) y un tiempo del proceso o tiempo de la ejecución (RunTime). El tamaño del proceso representa la cantidad de artículos que se fabrica en un caso determinado, el tiempo del proceso representa la cantidad de minutos que toma elaborar los artículos en el caso especificado. En los datos anteriores, el primer caso indica que para elaborar 175 artículos se requirió un tiempo de 195 minutos. El segundo caso indica que para elaborar 189 artículos se tomó un tiempo de 215 minutos. El último caso indica que, para elaborar 68 artículos, se requirió un tiempo de fabricación de 172 minutos. Para comenzar a trabajar con los datos deberá guardarlos en su directorio de trabajo. A continuación, importe los datos a R mediante la siguiente instrucción production <- read.table("production.txt", header = TRUE), puede observar el conjunto de datos en R al ejecutar la palabra production. Extraiga las columnas RunSize y RunTime del data frame production mediante la instrucción attach(production), es decir, de ahora en adelante podrá utilizar los vectores RunSize y RunTime en R. Realice el gráfico de dispersión de los datos al ejecutar la siguiente instrucción plot(RunSize, RunTime, xlab="Run Size", ylab = "Run Time").

Cada punto del gráfico de dispersión representa el tamaño del proceso y el tiempo del proceso de un caso determinado. Ajuste un modelo de regresión lineal simple a los datos en R y obtenga un resumen del modelo

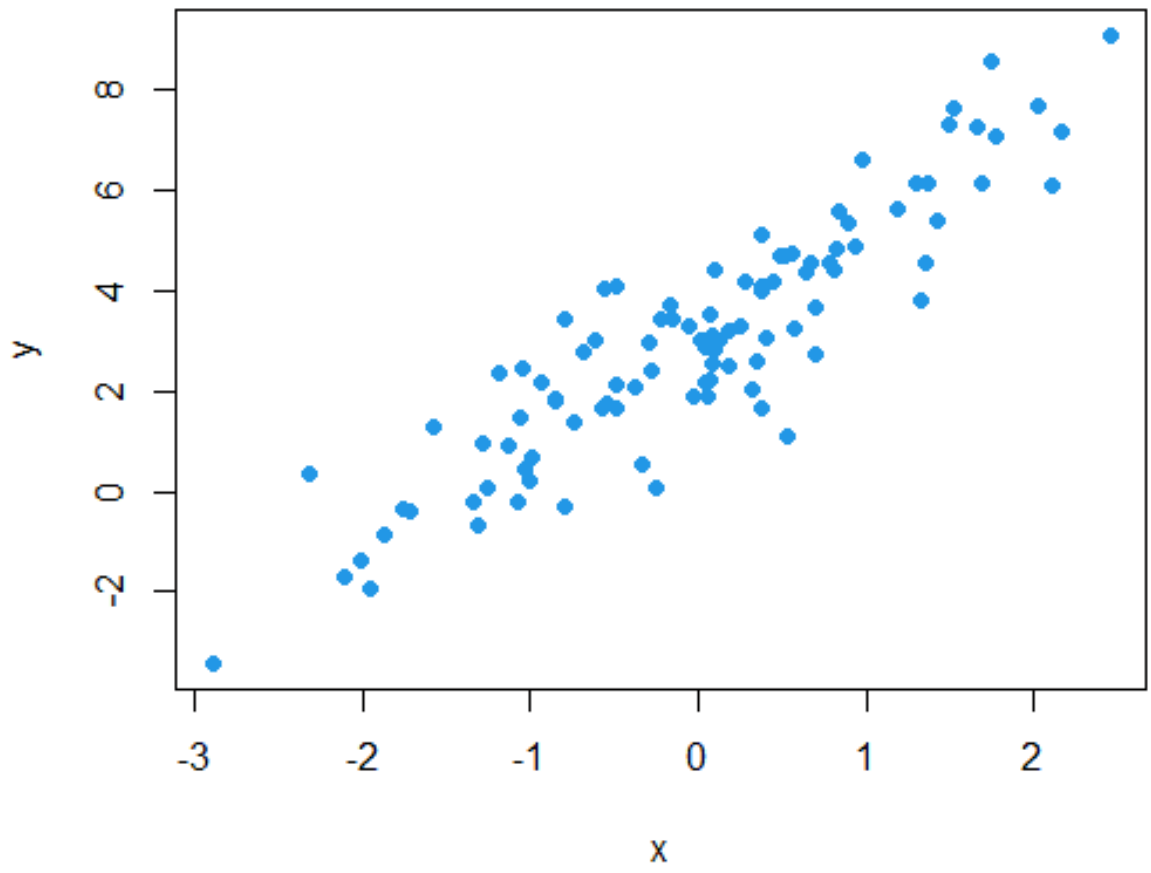

Figure 2: Ejemplo de regresion lineal

Case	RunTime	RunSize
1	195	175
2	215	189
3	243	344
4	162	88
5	185	114
6	231	338
7	234	271
8	166	173
9	253	284
10	196	277
11	220	337
12	168	58
13	207	146
14	225	277
15	169	123
16	215	227
17	147	63
18	230	337
19	208	146
20	172	68

Figure 3: tabla

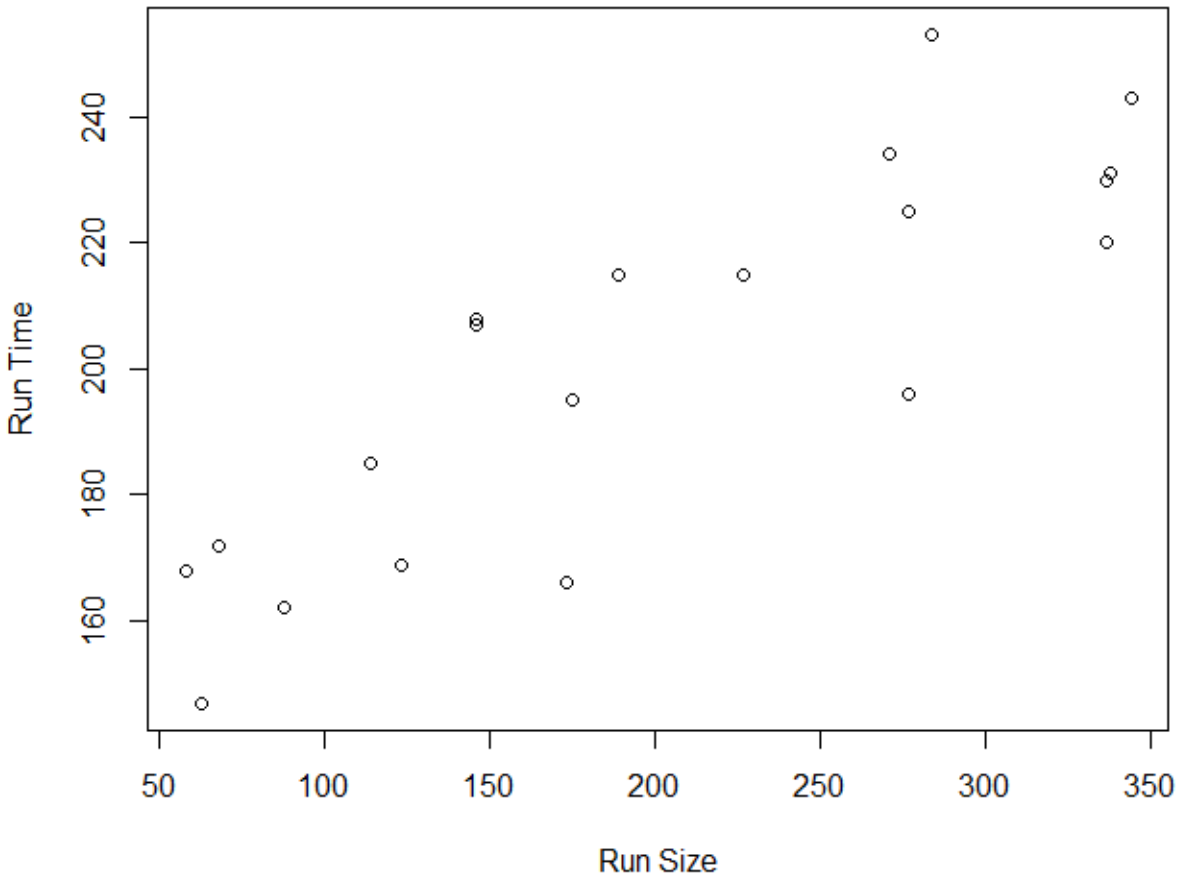

Figure 4: grafica

ajustado al ejecutar las siguientes dos instrucciones # Ajuste el modelo m1 <- lm(RunTime~RunSize) summary(m1)

```
Call:
lm(formula = RunTime ~ RunSize)

Residuals:
    Min      1Q  Median      3Q     Max
-28.597 -11.079   3.329   8.302  29.627

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.74770    8.32815   17.98 6.00e-13 ***
RunSize       0.25924    0.03714    6.98 1.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-squared:  0.7302,    Adjusted R-squared:  0.7152
F-statistic: 48.72 on 1 and 18 DF,  p-value: 1.615e-06
```

Figure 5: Imagen1

MAQUINAS DE VECTORES DE SOPORTE

Un enfoque para clasificación que se desarrolló en la comunidad de las ciencias computacionales en los años 90 y que ha crecido en popularidad desde entonces son las máquinas de vectores de soporte (MVS o SVM por sus siglas en inglés). Las MVS han mostrado un buen desempeño en una variedad de contextos, y frecuentemente se les considera como uno de los mejores clasificadores.

CLASIFICADOR DE MARGEN MAXIMO

Nuestro objetivo es desarrollar un clasificador basado en los datos de entrenamiento que clasificará una observación de prueba usando sus medidas características.

En un sentido, el hiperplano de margen máximo representa la línea media del bloque más ancho que podemos insertar entre las dos clases. Podemos calcular la distancia de cada observación de entrenamiento a un hiperplano de separación dado; la más pequeña de tales distancias es la distancia mínima de las observaciones al hiperplano y se conoce como el margen. El hiperplano de margen máximo es el hiperplano de separación para el cual el margen es el más grande-es decir, es el hiperplano que tiene la distancia mínima más lejana a las observaciones de entrenamiento-. En un espacio p-dimensional, un hiperplano es un subespacio plano de dimensión p-1 que no necesita pasar por el origen. En p dimensiones, un hiperplano se define por la ecuación

Podemos pensar al hiperplano como que divide el espacio p-dimensional en dos mitades. Por ejemplo, en dos dimensiones tenemos el hiperplano

El hiperplano de margen máximo es la solución al problema de optimización

sujeto a

La salida de R muestra entre muchas otras cosas el **intercepto estimado** $\hat{\beta}_0 = 149.7477$ y la **pendiente estimada** $\hat{\beta}_1 = 0.25924$ de la recta de regresión. Las fórmulas para obtener esos valores estimados a partir de los datos son las siguientes

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Donde los datos están dados por pares $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$.

Para graficar la recta de regresión estimada $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, ejecute la siguiente instrucción en R

```
# Graficar la recta estimada

abline(lsfit(RunSize, RunTime))
```

Figure 6: Imagen2

EL CASO NO SEPARABLE

Podemos extender el concepto de un hiperplano de separación para desarrollar un hiperplano que casi separa las clases usando lo que se conoce como un margen suave.

CLASIFICADOR DE VECTORES DE SOPORTE

La distancia de una observación al hiperplano puede considerarse como una medida de nuestra confianza de que la observación se clasifica correctamente. Podemos estar dispuestos a considerar un clasificador basado en un hiperplano que no separe perfectamente las dos clases, con el interés de: * Mayor robustez a observaciones individuales, y Mejor clasificación de la mayoría de las observaciones de prueba.

Es decir, podría valer la pena clasificar mal unas pocas observaciones de entrenamiento para hacer un mejor trabajo al clasificar las observaciones restantes. Un hiperplano que casi separa las clases es la solución al problema de optimización.

Sujeto a:

M es el ancho del margen; buscamos hacer esta cantidad tan grande como sea posible. Una vez que hemos resuelto el problema de optimización, clasificamos una observación de prueba x* como antes, al simplemente determinar de que lado del hiperplano se encuentra. Es decir, clasificamos la observación de prueba basados en el signo de

Conforme el presupuesto C se incrementa, nos volvemos más tolerantes con respecto a las violaciones al margen, y así el margen se hará ancho. Por otro lado, cuando C decrece, nos volvemos menos tolerantes a las violaciones al margen y así el margen se hace angosto. En la práctica, C es tratada como un parámetro que generalmente se elige por medio de validación-cruzada. **Las observaciones que se encuentran directamente sobre los márgenes o del lado incorrecto del margen considerando su clase, se conocen como vectores de soporte. Clasificación con frontera de decisión no lineal.**

En el caso del clasificador de vectores de soporte, podemos tratar el problema de posibles fronteras no-lineales entre clases al ampliar el espacio de características usando funciones polinomiales cuadráticas, cúbicas, o

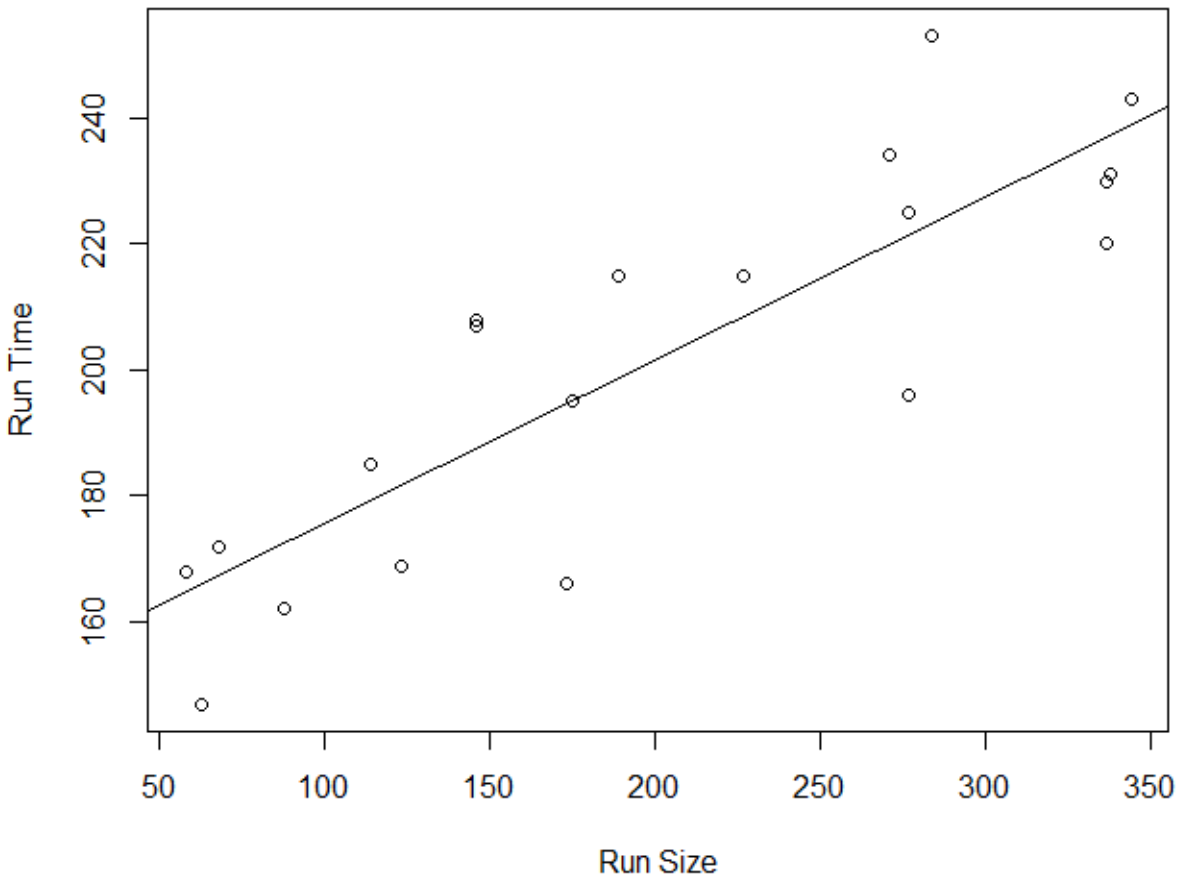

Figure 7: Imagen3

Los residuales del ajuste están definidos por $\hat{e}_i = y_i - \hat{y}_i$, es decir, a cada valor y_i que tenemos, le restamos el valor estimado $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Para obtener los residuales del ajuste en R ejecute la siguiente instrucción m1$residuals

Figure 8: Imagen4

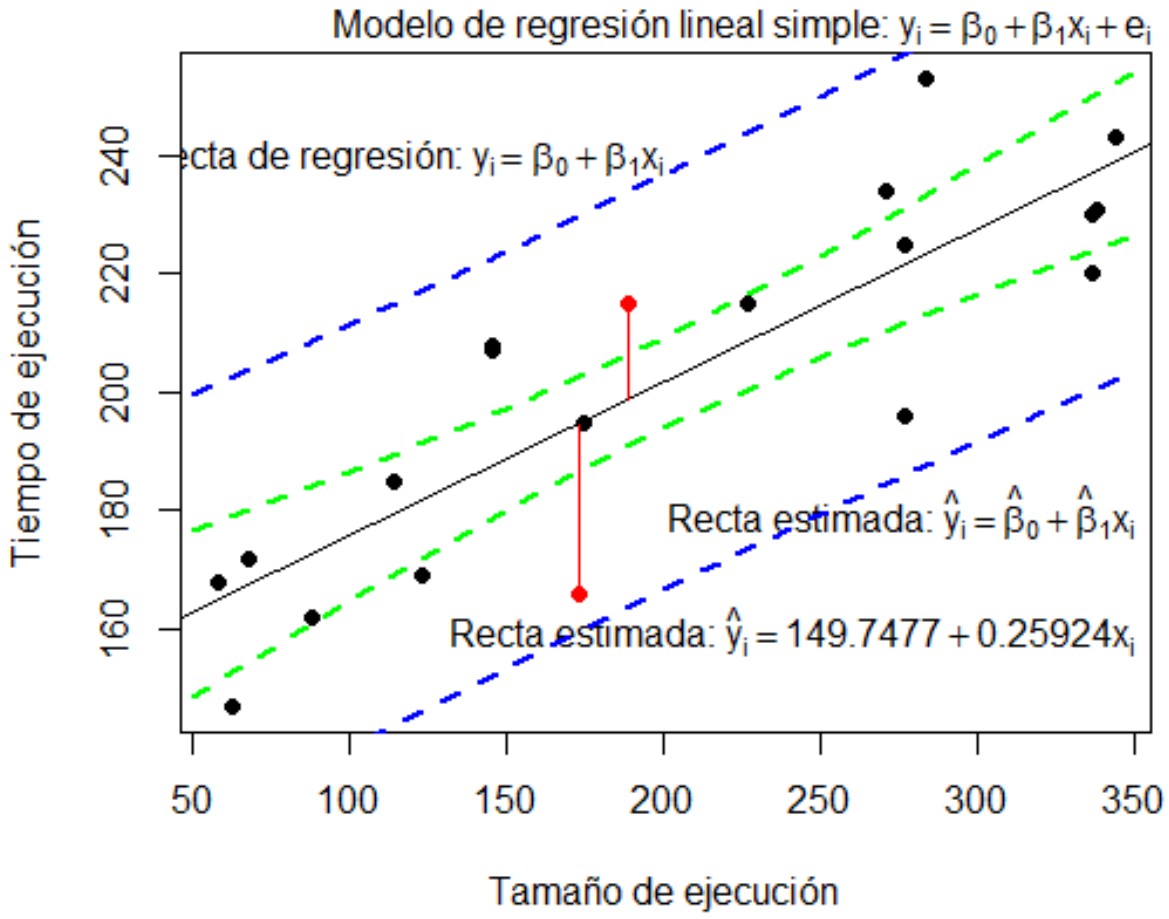

Figure 9: Imagen5

Obtenga la media de los residuales mediante mean(m1$residuals)

Obtenga una estimación de la varianza de los errores mediante sum(m1$residuals^2)/18

Figure 10: Imagen6

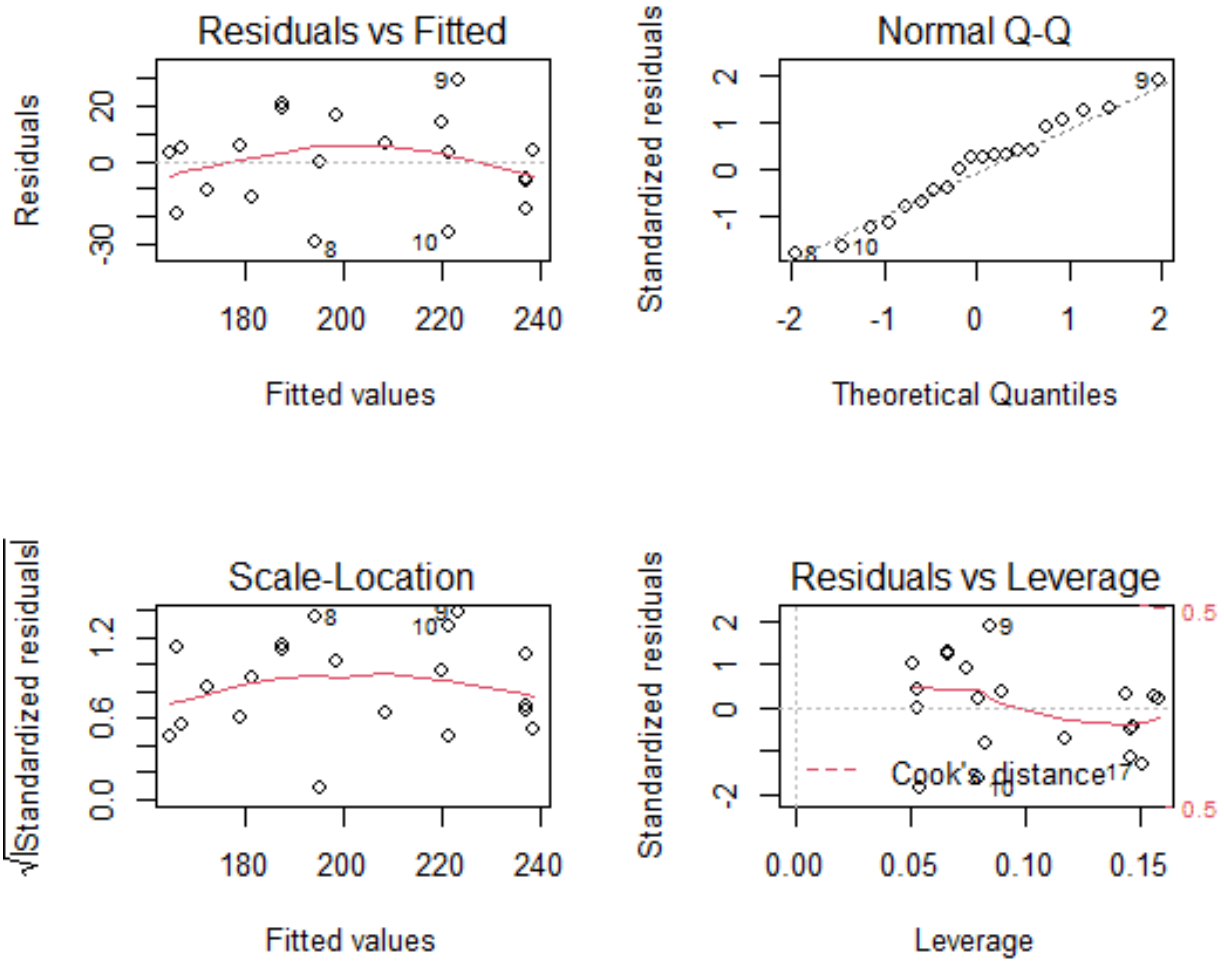

Figure 11: Imagen7

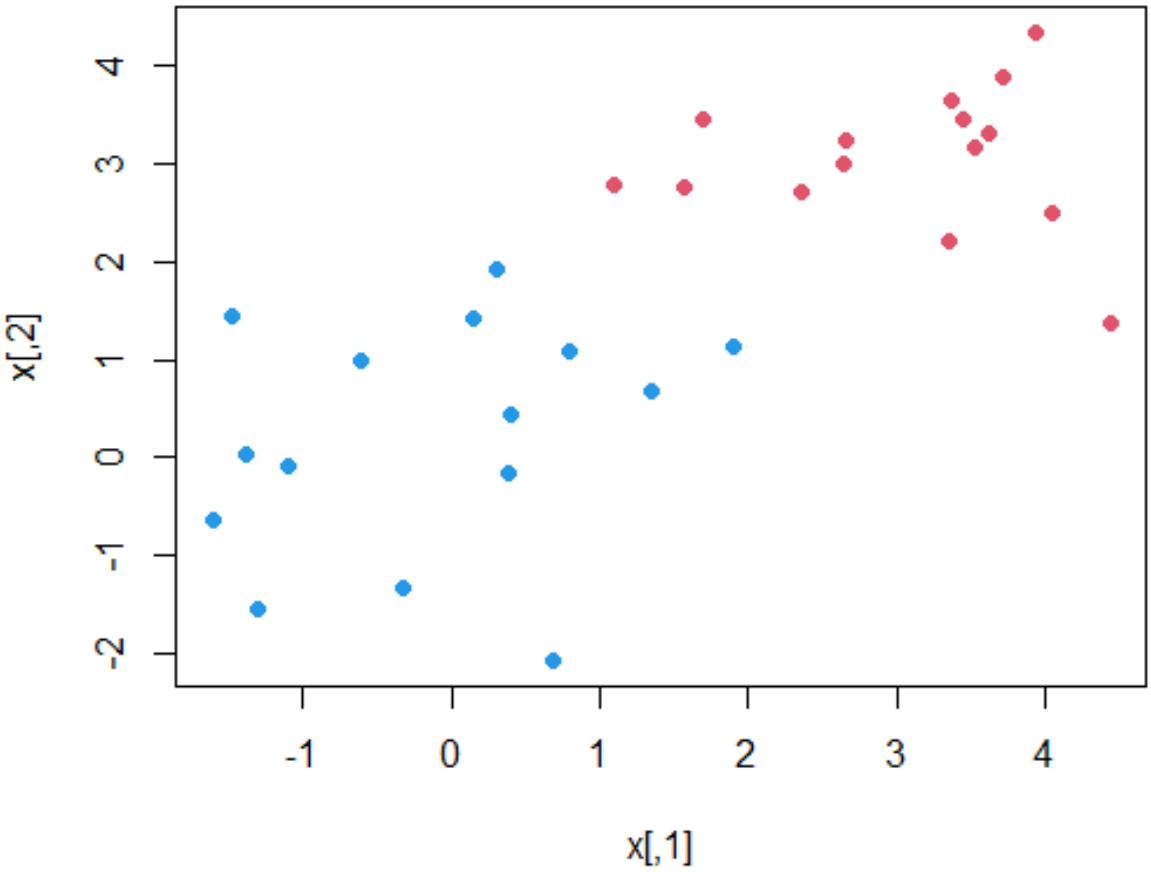

Figure 12: Imagen8

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

Figure 13: Imagen9

$$1 + 2X_1 + 3X_2 = 0$$

Figure 14: Imagen10

$$1 + 2X_1 + 3X_2 = 0$$

Figure 15: Imagen11

Si $\beta_0, \beta_1, \cdots, \beta_p$ son los coeficientes del hiperplano de margen máximo, entonces el clasificador de margen máximo clasifica la observación de prueba x^* basado en el signo de

Figure 16: Imagen12

$$1 + 2X_1 + 3X_2 = 0$$

Figure 17: Imagen13

Si $\beta_0, \beta_1, \cdots, \beta_p$ son los coeficientes del hiperplano de margen máximo, entonces el clasificador de margen máximo clasifica la observación de prueba x^* basado en el signo de

Figure 18: Imagen14

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$$

Figure 19: Imagen15

$$\max_{\beta_0,\beta_1,\cdots,\beta_p,M} M$$

Figure 20: Imagen16

$$\sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M \quad \forall i = 1, \cdots, n.$$

Figure 21: Imagen17

incluso de orden superior de los predictores. Por ejemplo, en vez de ajustar un clasificador de vectores de soporte usando p características

podríamos ajustar un clasificador de vectores de soporte usando 2p características

Entonces el problema de optimización

sujeto a

Se convertiría en:

sujeto a

No es difícil ver que hay muchas maneras de ampliar el espacio de características, y que a menos que seamos cuidadosos, podríamos terminar con un número enorme de características. Entonces los cálculos serían inmanejables.

LA MAQUINA DE VECTORES DE SOPORTE

La máquina de vectores de soporte (SVM por sus siglas en inglés) es una extensión del clasificador de vectores de soporte que resulta de ampliar el espacio de características de una manera específica, usando kernels. Podemos querer ampliar nuestro espacio de características para acomodar una frontera no-lineal entre las clases. El enfoque del kernel que describimos aquí es simplemente un enfoque computacional eficiente para llevar a cabo esta idea.

PUEDE DEMOSTRARSE QUE **1**. El clasificador de vectores de soporte lineal se puede representar como:

donde hay n parametros i,$i = 1, ..., n$, uno por observacion de entrenamiento.

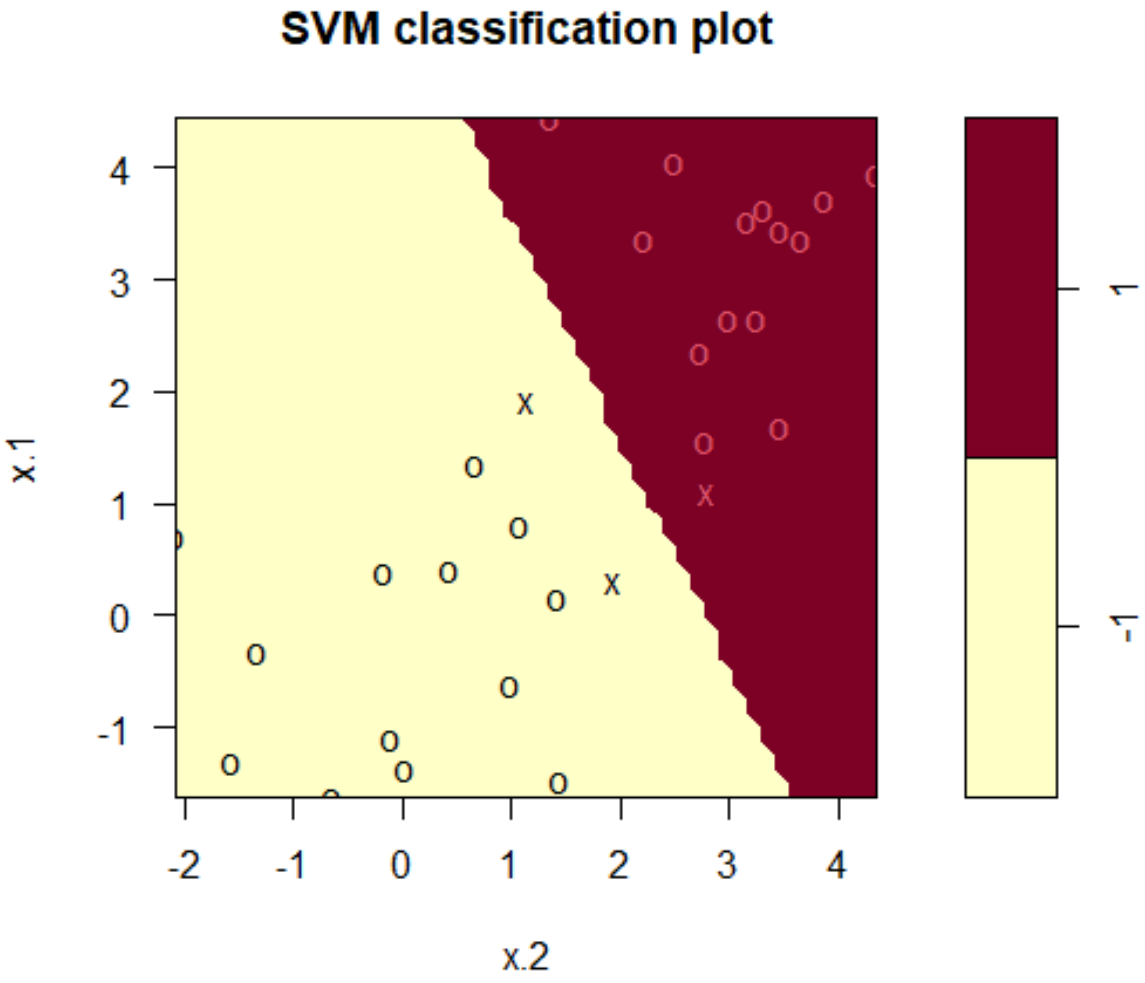

Figure 22: Imagen18

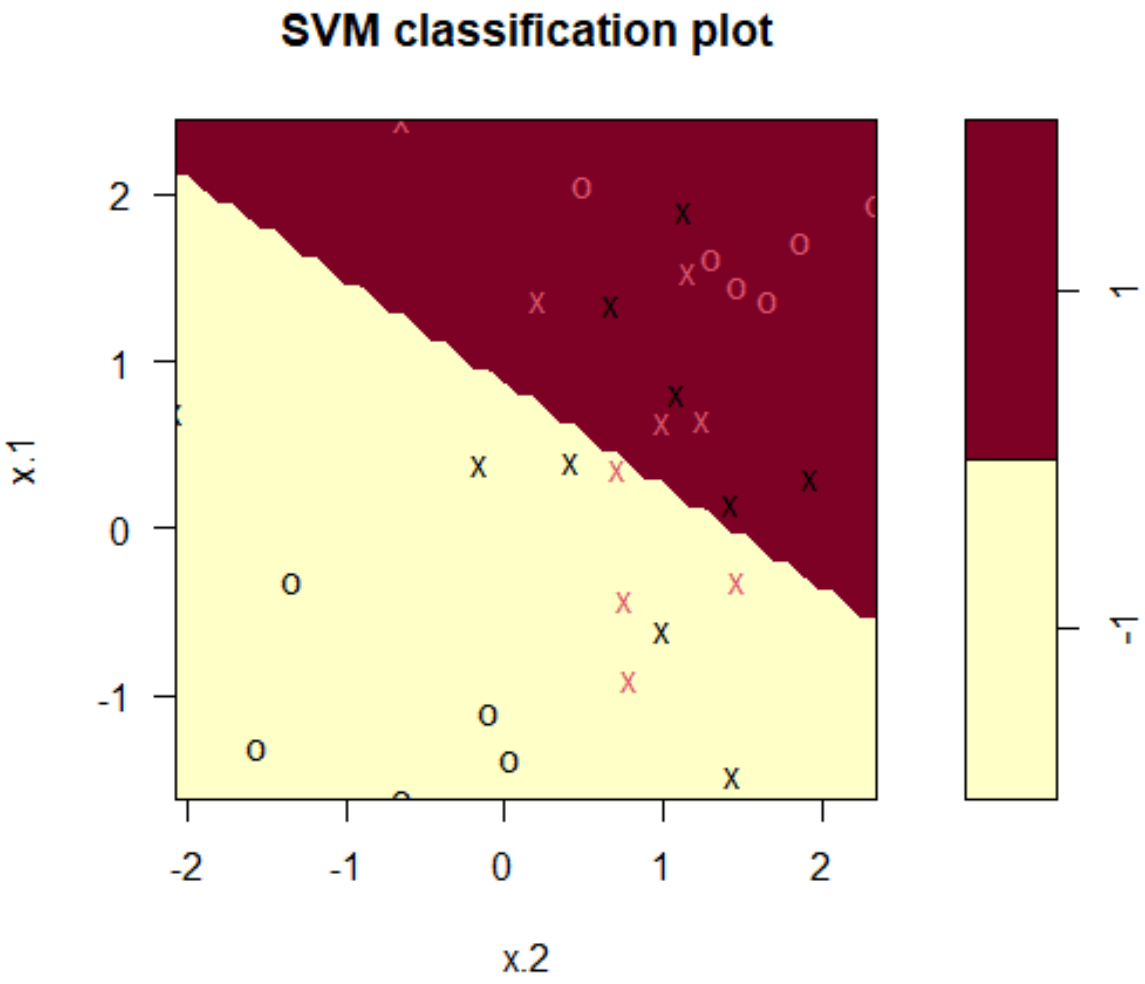

Figure 23: Imagen19

$$\max_{\beta_0,\beta_1,\cdots,\beta_p,\varepsilon_1,\cdots,\varepsilon_n,M} M$$

Figure 24: Imagen20

$$\sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i),$$

$$\varepsilon_i \geq 0, \quad \sum_{i=1}^{n} \varepsilon_i \leq C,$$

Figure 25: Imagen21

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$$

Figure 26: Imagen22

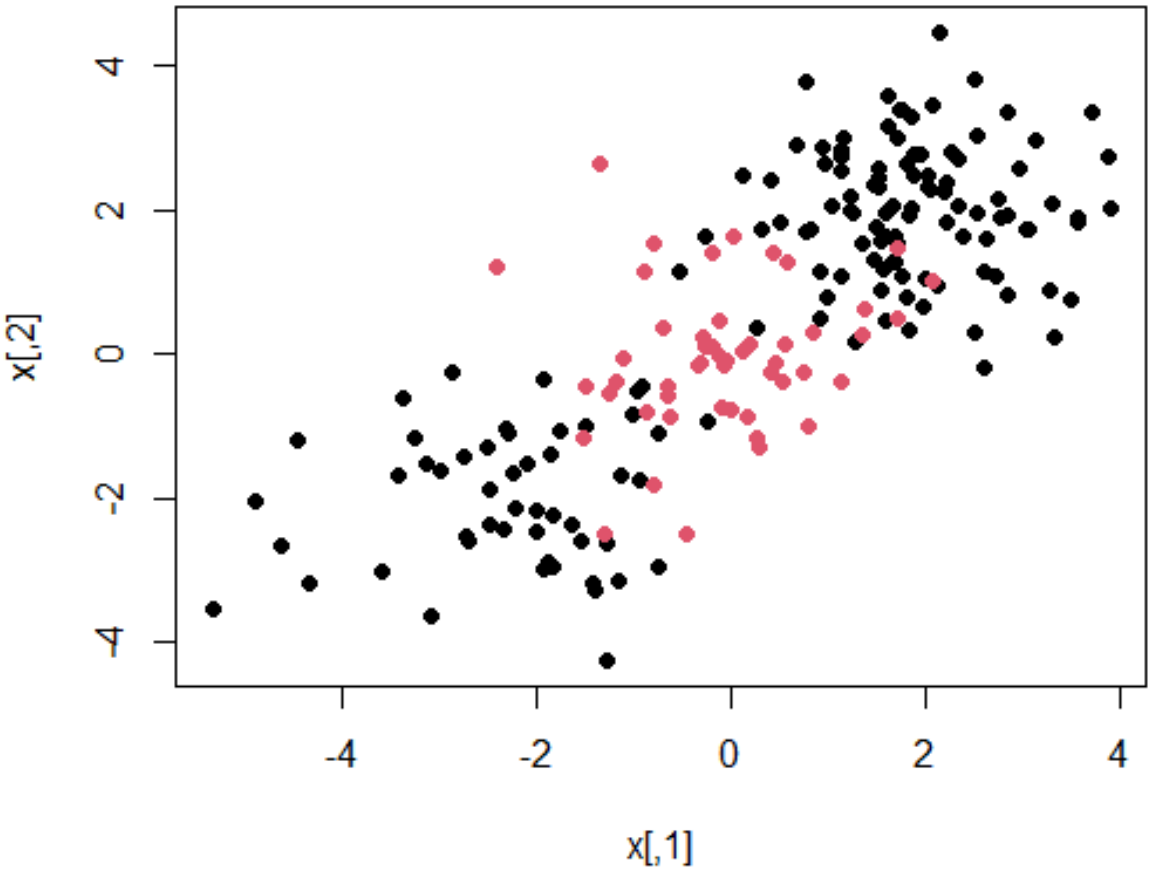

Figure 27: Imagen23

$$X_1, X_2, \cdots, X_p,$$

Figure 28: Imagen

$$X_1, X_1^2, X_2, X_2^2, \cdots, X_p, X_p^2.$$

Figure 29: Imagen

$$\max_{\beta_0,\beta_1,\cdots,\beta_p,\varepsilon_1,\cdots,\varepsilon_n,M} M$$

Figure 30: Imagen

$$\sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i),$$

$$\varepsilon_i \geq 0, \quad \sum_{i=1}^{n} \varepsilon_i \leq C,$$

Figure 31: Imagen

$$\max_{\beta_0,\beta_{11},\beta_{12},\cdots,\beta_{p1},\beta_{p2},\varepsilon_1,\cdots,\varepsilon_n,M} M$$

Figure 32: Imagen

$$\sum_{j=1}^{p}\sum_{k=1}^{2}\beta_{jk}^2 = 1,$$

$$y_i(\beta_0 + \sum_{j=1}^{p}\beta_{j1}x_{ij} + \sum_{j=1}^{p}\beta_{j2}x_{ij}^2) \geq M(1-\varepsilon_i),$$

$$\varepsilon_i \geq 0, \quad \sum_{i=1}^{n}\varepsilon_i \leq C.$$

Figure 33: Imagen

El producto interno de dos observaciones x_i , $x_{i'}$ está dado por

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}.$$

Figure 34: Imagen

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle,$$

Figure 35: Imagen

donde hay n parámetros α_i , $i = 1, \cdots, n,$ uno por observación de entrenamiento.

2. Para estimar los parámetros $\alpha_1, \cdots, \alpha_n$ y β_0 , todo lo que necesitamos son los $\binom{n}{2}$ productos internos $\langle x_i, x_{i'} \rangle$ entre todos los pares de observaciones de entrenamiento.

Resulta que α_i es diferente de cero sólo para los vectores de soporte en la solución-es decir, si una observación de entrenamiento no es un vector de soporte, entonces su α_i es igual a cero-.

Si S es la colección de índices de estos puntos de soporte, podemos re-escribir cualquier función de solución como

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle$$

Figure 36: Imagen

KERNEL

Un kernel es la funcion que cuantifica la similaridad de dos observaciones.

KERNEL LINEAL

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j},$$

Figure 37: Imagen

KERNEL POLINOMIAL

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d.$$

Figure 38: Imagen

Cuando el clasificador de vectores de soporte se combina con un kernel no-lineal, el clasificador que resulta se conoce como una máquina de vectores de soporte. En este caso la función tiene la forma

KERNEL RADIAL

Comportamiento local del kernel radial

CLASIFICACIÓN CON FRONTERA DE DECISION NO LINEAL

¿Cuál es la ventaja de usar un kernel en lugar de simplemente ampliar el espacio de características usando funciones de las características originales?

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i).$$

Figure 39: Imagen

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

Figure 40: Imagen

$$K(x^*, x_h) = exp(-\gamma \sum_{j=1}^{p} (x_j^* - x_{hj})^2).$$

$$f(x^*) = \beta_0 + \sum_{i \in S} \alpha_i K(x^*, x_i).$$

Figure 41: Imagen

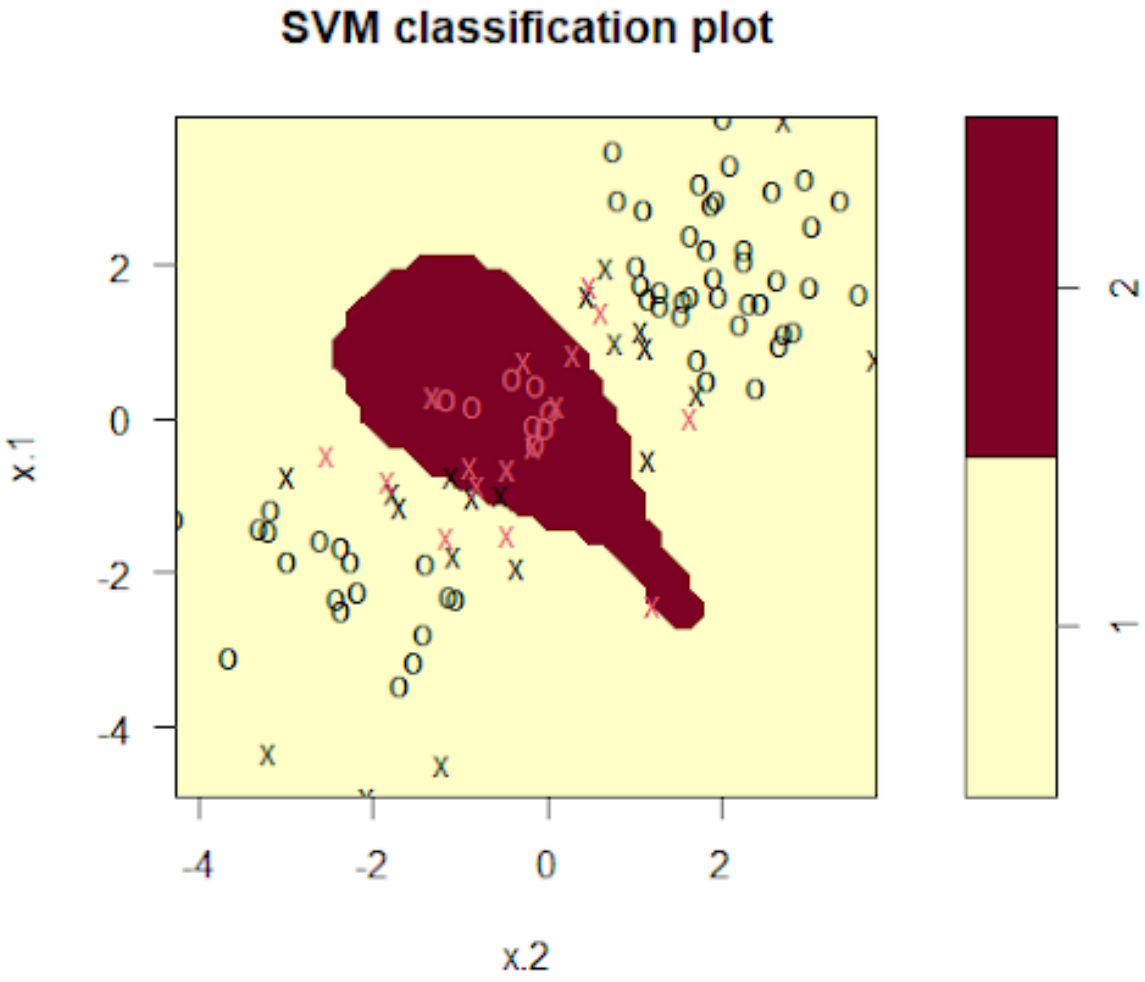

Figure 42: Imagen

Sólo necesitamos calcular $K(x_i, x_{i'})$ para todos los pares distintos i, i'.

Figure 43: Imagen