Práctica 2: Regresión Logística

Esta práctica aplicar el método de regresión logística sobre unos datos de entrenamiento previamente dados.

La regresión logística nos sirve para clasificar unos datos determinados en varias clases, para una vez encontrada la "función de clasificación", poder predecir en qué clase irán nuevos datos que todavía no han aparecido.

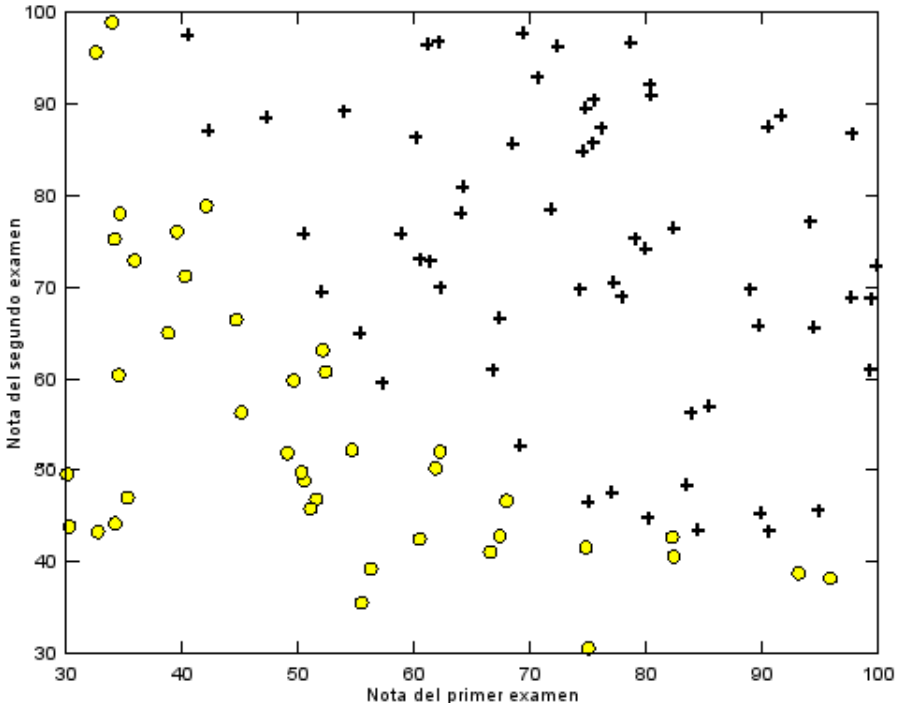
En esta práctica conseguiremos aproximar una función sigmoide que nos sirva para determinar la probabilidad de que un alumno sea admitido por una universidad en base a la nota de 2 exámenes, y determinar si un chip pasará o no el control de calidad en base a 2 test.

Hemos empleado 2 métodos de regresión:

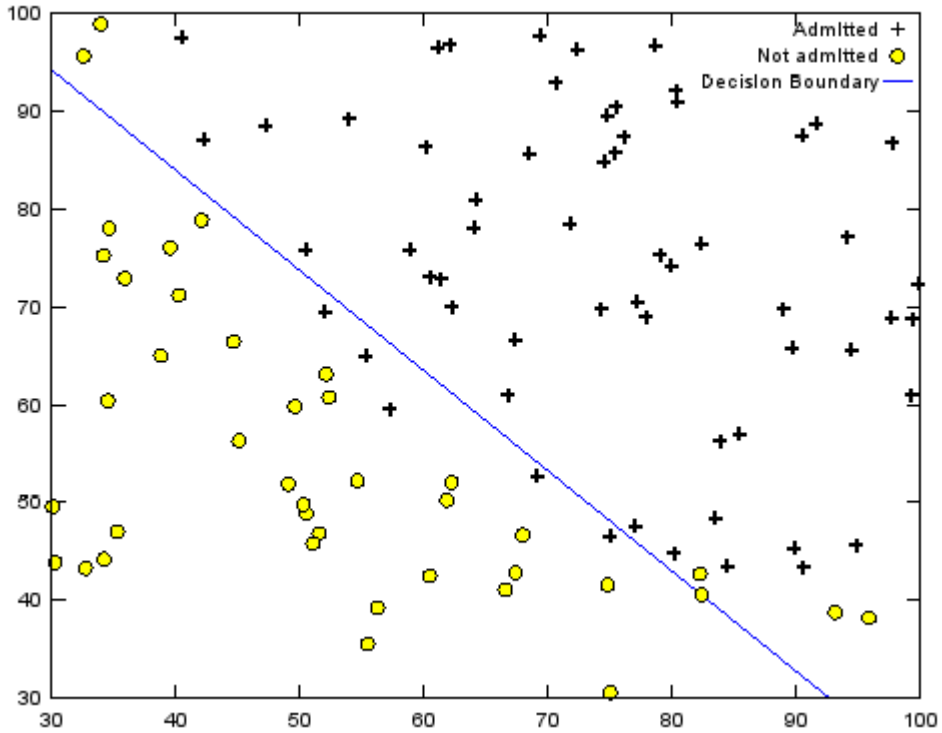
Regresión logística lineal

En el primer caso podemos aplicar la regresión logística porque existe una línea que puede separar los casos de ejemplo positivos de los negativos.

Como se puede comprobar, los casos de estudio están relativamente separados los positivos de los negativos. Y aunque no se separan exactamente por una línea recta, sí que está cercano a ser una recta.



Mediante el código que se añade al final del fichero, se ha calculado cual es la función que diferencia los positivos de los negativos, pudiendo dibujar la recta que los separa.



Como se aprecia en la imagen, el acierto no es del 100%. Esto es porque se hace una aproximación lineal, y por lo tanto no se pueden abarcar todos los ejemplos de entrenamiento.

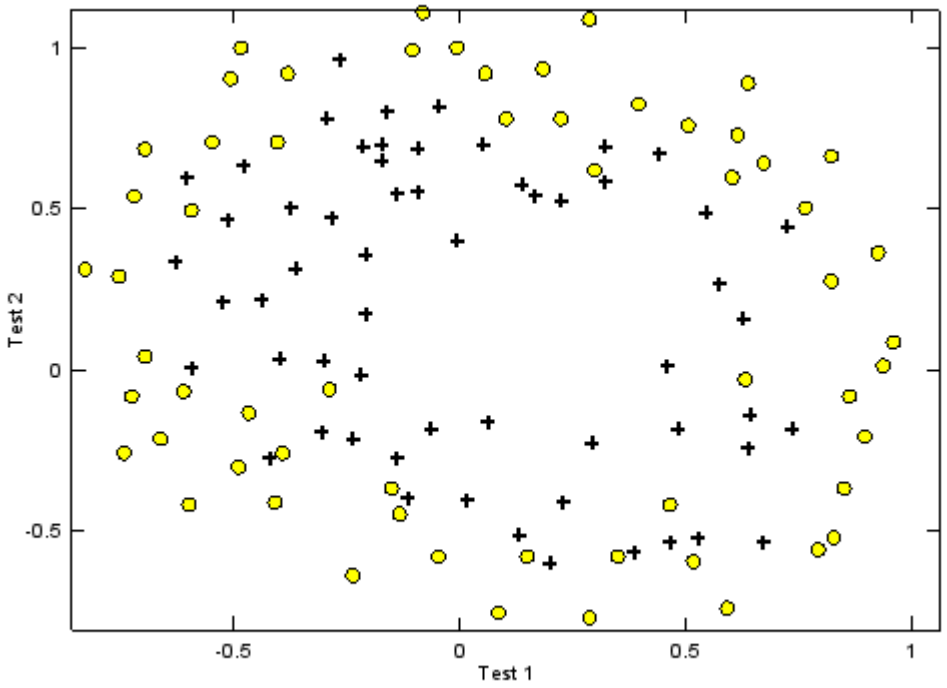
Además, en el código, hemos calculado el porcentaje de aciertos de la función, determinado como la cantidad de ejemplos que la función ha claculado correctamente. En este caso, nos da una probabilidad de acierto de 0.89, que es una cifra bastante elevada.

Regresión logística regularizada

Con estos segundos datos de entrenamiento hemos utilizado el método de regresión logística regularizada porque se parte de unos datos de entrenamiento que sería imposible dividir con una línea recta.

En la regresión logística regularizada lo que hacemos es intentar buscar una función con más términos (una elipse) para representar funciones más complejas, ya que con una línea recta no podríamos determinar una función representativa. Por lo tanto nunca nos ajustaríamos a los datos reales, y nunca podríamos hacer una predicción futura.

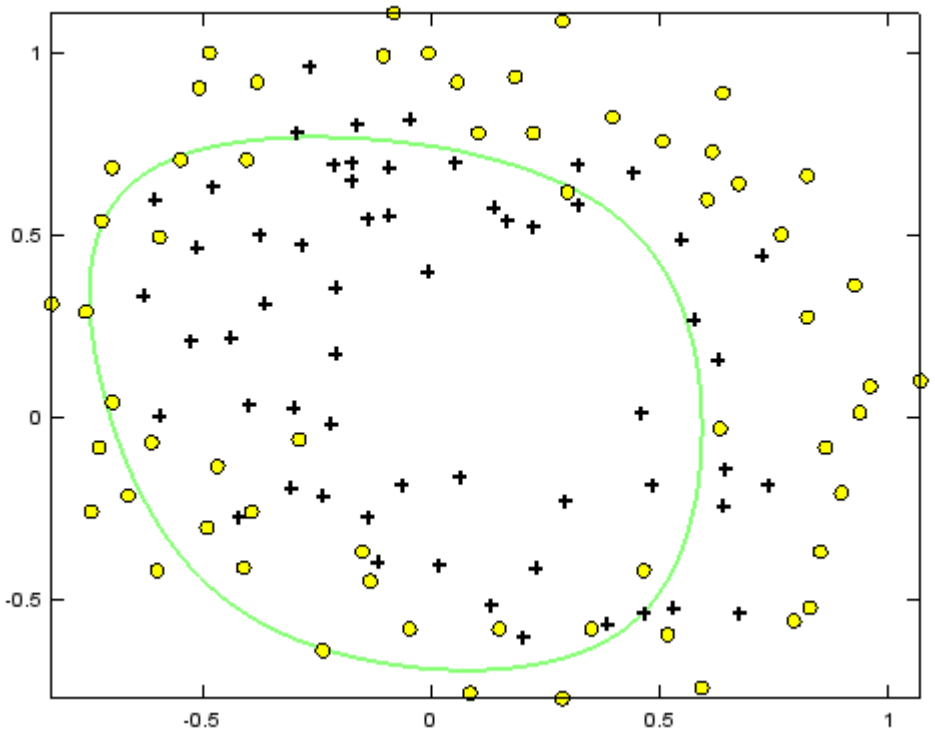
Lo primero ha sido analizar los datos, donde se ve claramente lo que hemos explicado acerca de la división de los mismos.



En este caso lo que debemos hacer es algo similar a la regresión logística, pero con la diferencia de que “extendemos” el número de atributos numerosas veces, hasta 28 atributos, para poder crear funciones más complejas.

Ya con estos atributos, realizamos cálculos similares para hallar la función de coste, que iremos optimizando mediante el gradiente.

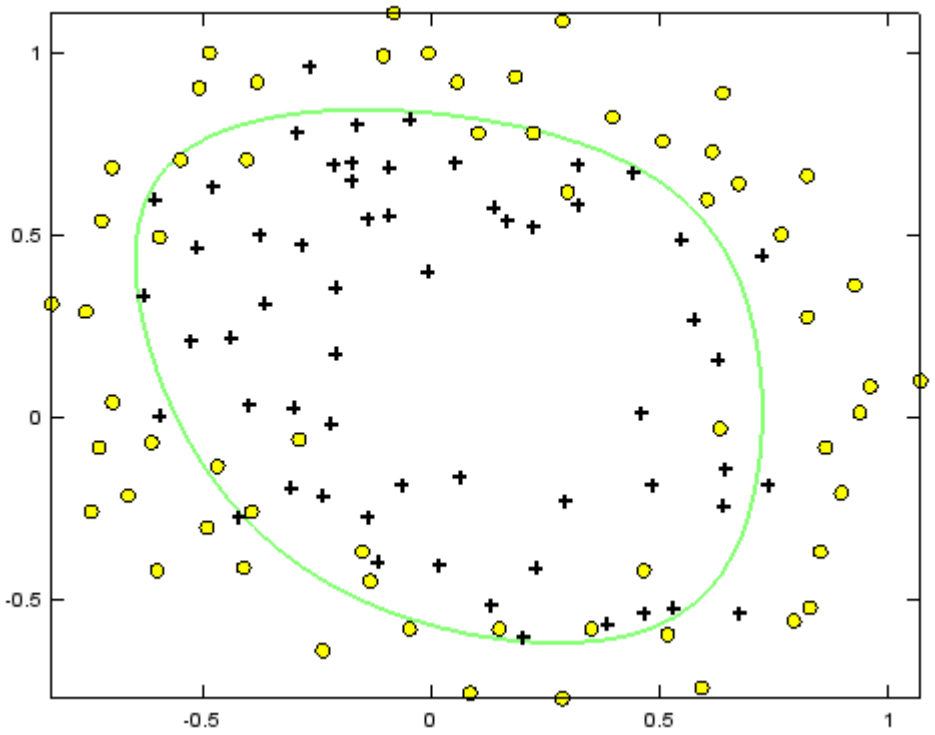
Con un valor de $\lambda = 1$ obtenemos una función como la siguiente:



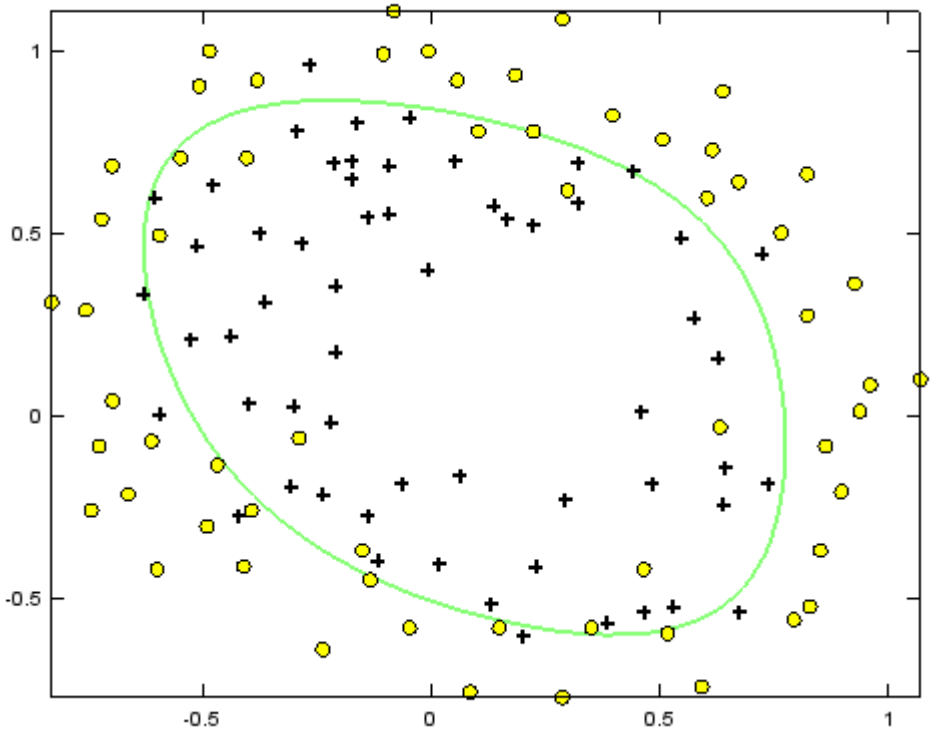
Pero si vamos modificando el valor de landa a valores más pequeños, vamos ajustando más la función, obteniendo una mayor probabilidad de acierto en los ejemplos de entrenamiento.

Esto, si se hace demasiado puede ser contraproducente, pues cuanto más lo ajustemos a los ejemplos de entrenamiento, menor será la capacidad de predicción para nuevos valores.

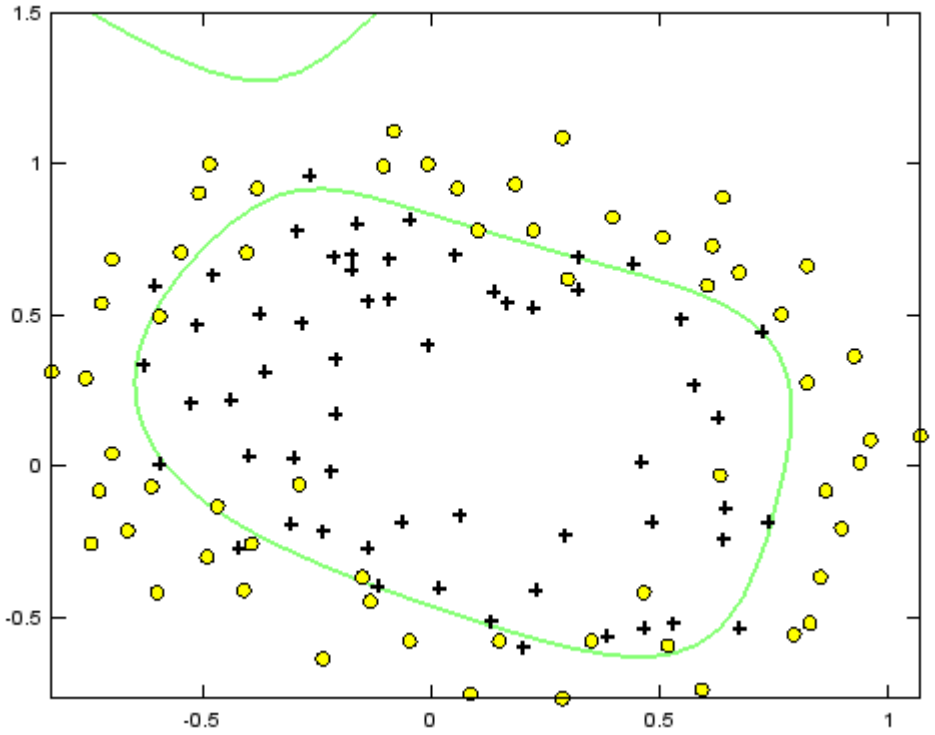
λ = 0,1



λ = 0,01



λ = 0,001



λ = 0,0001