Universidad Nacional Autónoma de México
Facultad de Ingeniería

Programa de Tecnología en Cómputo

MatLab

24 de Mayo del 2019

Modelos de Machine Learning: Regresión Multivariable y Logística

Alumno

Garrido Sánchez Samuel Arturo

González Hernández Adriana

Maceda Patricio Fernando

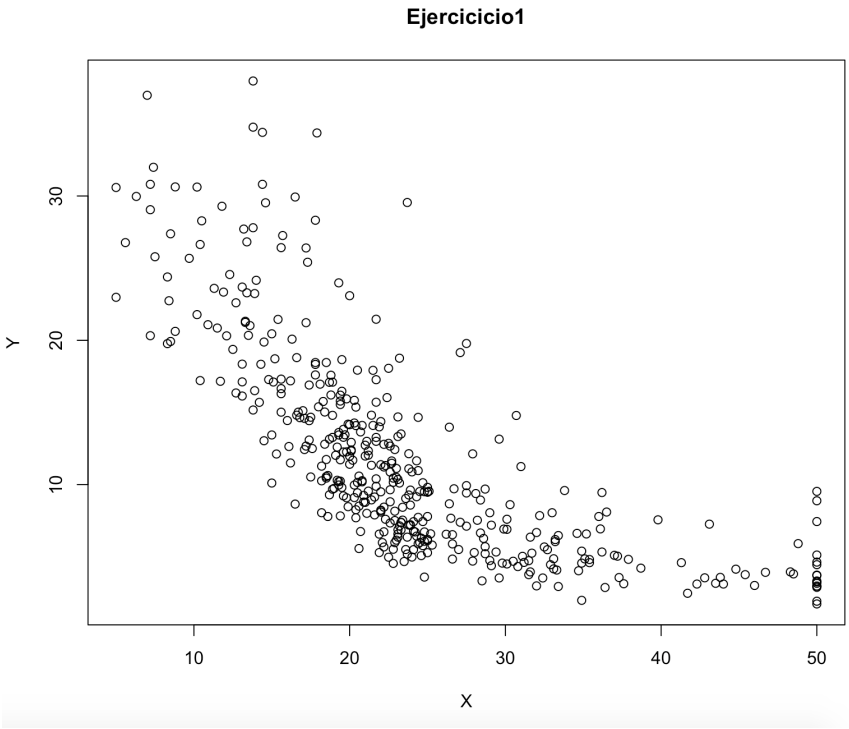
Curso de selección
Ciudad Universitaria, Coyoacán, CDMX

1. Introducción

Cuando hablamos del reconocimiento visual de algunas páginas en internet, de la publicidad contextual en función de los hábitos de navegación o las recomendaciones de sitios de acuerdo a la segmentación o pertenencia a determinados grupos, hablamos de información que se recoge gracias al historial de navegación y es así como podemos definir el aprendizaje de las máquinas o Machine Learning. Cada ser humano, segundo a segundo, es una fuente de generación de datos sobre los intereses, valores y preferencias de consumo que se registran a través de las redes sociales como Instagram, Facebook, WhatsApp, e-mail, LinkedIn, Twitter, Pinterest, entre otros. La mercadotecnia electrónica, la web, todo tipo de transacciones bancarias y comerciales o entre máquina y máquina como Wi-Fi, Bluetooth, GPS o la navegación por inter- net, arrojan los datos masivos que gracias al pro- cesamiento a través de aplicaciones informáticas, permite la manipulación y gestión predictiva.

2. Primer problema: Regresión Lineal Multivariable

Como parte de la resolución del primer ejercicio hay que saber el tipo de datos que estamos tratando y qué deseamos obtener de ellos. Como primer lugar tenemos nuestro DataSet, que consiste de un feature y un target, por lo que la regresión será lineal simple.



Como elementos en éste problema se trata con las siguientes funciones:

- **Hipótesis:** Tenemos que tener una función hipótesis la cuál será la función matemática que pueda precedir la tendencia de un nuevo dato cuando se le ingrese otro dato. La función hipótesis en éste caso fue compuesta por un vector que contiene 2 coeficientes θ_1 y θ_2. Estos corresponderán a la pendiente y la ordenada al origen en éste modelo. Si queremos modificar nuestra función hipótesis asignados más θ_i en el archivo de costoComputacional.m que contiene a la hipótesis la cual el vector está compuesto únicamente por X*theta.

- **Costo:** Una vez que poseemos nuestra hipótesis vamos a determinar el costo que supone tener una predicción errónea del modelo, para esto recurrimos al error cuadrático medio que se encarga de ver qué tan alejado se encuentran los features de la función matemática, ésta podemos representarla o decir que su tranformación a un campo distinto se convierte a una parábola en el cuál mediante la ecuación (en el caso lineal):

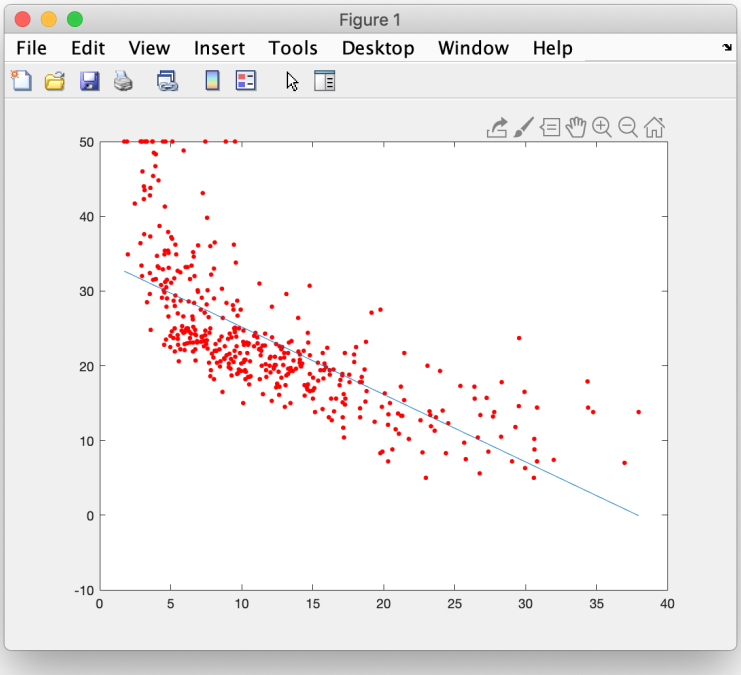
$$\sigma_1 = \sigma_1 - \alpha\sigma_1$$

  Esto podemos verlo como vamos quitando valores a sigma (costo) hasta que éste éste sea 0 para obtener el mínimo global al ser una función convexa.

- **Entrenamiento:** Para lograr el objetivo anterior es necesario tener una función que nos realize las iteraciones en el programa por lo que recurrimos al descenso gradiente para poder lograrlo. Para logra ésto hay que recordar que la función descenso gradiente está dada por:

$$\theta = \theta - \alpha\frac{\partial f}{\partial \theta}$$

  Dado que nos encontramos en un sistema simple la derivada estará dado por solo theta. Así que de manera banal podemos describirlo como que ubicamos nuestra recta aleatoriamente y la vamos moviendo hasta que la derivada parcial en ese punto sea 0, o como geométricamente se puede decir: Cuando la pendiente es 0 ya estamos en el mínimo así que en teoría vamos disminuyendo la pendiente hasta que $\frac{\partial f}{\partial \theta}$ sea despreciable y terminamos con el entrenamiento.



- **Prueba:** Para poder garantizar que nuestro modelo ha llegado a un punto de predicción cercano al valor patrón, sustituimos en la hipótesis para observar si se encuentra en o cerca de la recta del modelo.

Archivos utilizados:

- **Ejercicio1**
- CostoComputacional
- GradienteDescenso

Para éste ejercicio podemos dar por conclusión que éstas iteraciones nos permiten inferir valores futuros dado datos actuales. ¿Por qué rayos entonces no hacemos una regresión lineal común con todo el dataSet? ¿Más preciso? Bueno el hecho es que éstos DataSet son de prueba y no está ligero soportar Tablas de millares de datos. Por ello esta iteración garantiza que creemos un modelo aproximado y que con la llegada de nuevos features no sea necesario repasar todos lo datos hasta el momento. Si aplicamos regresión lineal normal tendríamos además complejidades arriba del n^2 por lo que en la práctica resulta ineficiente.

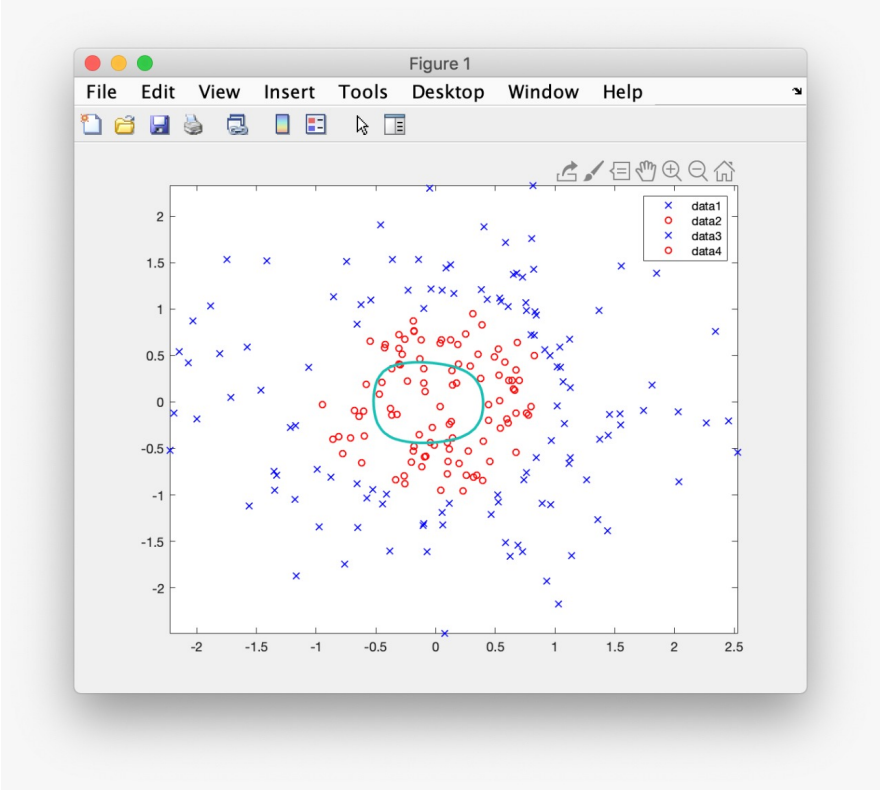
3. Segundo problema: Regresión logística

El segundo problema resultó algo más complejo que el primero, ya que tenemos que determina en qué punto clasificamos de un tipo a un dato. Vale recalcar que éste algoritmo es de clasificación que es tal que su codominio es discreto. En éste caso, el dataset nos ofrece 2 features y nos devuelve un target que es binario. Ahora con éstos graficamos y podemos a vista saber que se ubicará la frontera de decisión cerca de la ecuación:

$$x^2 + y^2 = 1$$

, una circunferencia de radio 1.

- **Hipótesis:** Para la función hipótesis necesitamos antes de la función MapeoDelFeature al tener una función de éste tipo. Dentro podemos asignar el grado que tendrá theta debido a que deben haber 2 valores cuadrados para modelar un círculo. DE la matriz modificada X establecemos nuestro coeficiente de aprendizaje = 1.
- **Función sigmoide:** En éstos casos de clasificación es utilizada la función sigmoide para asignar a todo real un lugar dentro del rango $(0, 1)$. Expresamos nuestra función sigmoide en FuncionSigmoide.m
$$c = \frac{1}{1 + e^{-x}}$$
- **Función costo:** Para determinar el costo que podemos expresarlo como un recorrido de la función hipótesis en el campo de los features, debe ser tal que cubra a la gran mayoría de los aceptados y establezca un límite entre las clasificaciones. Dentro de nuestra función fue modificada para que pueda aceptar 2 thetas de la hipótesis como cuadráticas de manera que J costo seguirá establecido como costo y grad como el vector establecido para corregir la posición del lugar geométrico.
- **Entrenamiento:** Para el entrenamiento es necesario que tengamos en cuenta un optimset que es una función dentro de MatLab que nos permite realizar iteraciones de manera de clasificación dentro del Toolkit de Optimización. Ahora es necesario establecer la razón de cambio para mover la frontera de clasificación, que estará dado una función de tiempo fminunc(@(t)(CostoLogReg())). Dada a ésta tendremos el movimiento de la frontera de clasificación. Además de plotDecisionBoundary.



- **Prueba:** El reajuste nos determinará si se encuentra dentro o fuera del círculo que nos representa la frontera de decisión

Archivos utilizados:

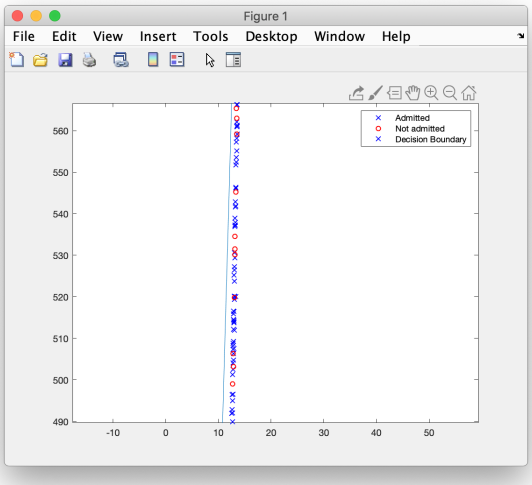
- **Logistica**
- MapeoDelaFuncion
- CostoLogisticaReg
- FuncionSigmoide
- plotDecision
- impresionN

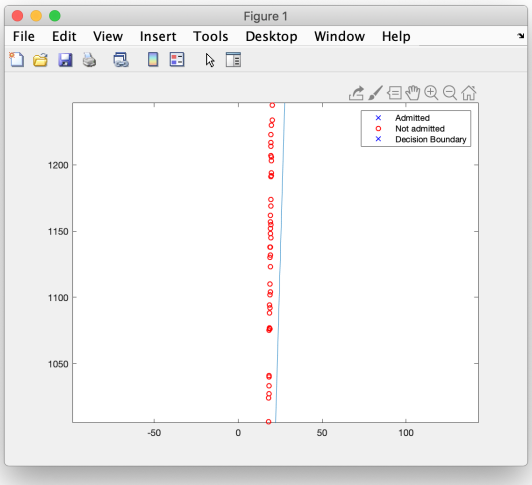
Dentro de éste ejercicio pudimos establecer un entrenamiento para determinar la ubicación del lugar geométrico.

4. Tercer problema: Regresión logística

En esta parte se nos entregó un archivo csv que contenía 31 características de un tumor de diversos pacientes y determinaba si eran malignos o no. Se nos permitió hacerlo con los targets que considerábamos necesarios. Éste ejercicio también era una regresión logística por lo que usamos el procesamiento del punto anterior con ciertas diferencias en el pre procesamiento. Fue utilizado las características de: Radio del tumor y la concavidad del mismo. La modificación se encuentra en la

función de costo, ya que en éste caso solamente necesitamos una línea divisora y que en el modelo anterior necesitábamos una curva cerrada.
En este caso tenemos la misma cantidad de features pero esta ocasión no tendremos una función de que contenga dobles thetas cuadradas por lo que usaremos la función CostoLogistica normal.





5. Conclusiones

Fue un proyecto introductorio a Machine Learning donde se tuvieron que manejar los conceptos de regresión lineal y logística multivariada, función de hipótesis, gradiente de descenso, función de costo, entre otros. Aunque todo el sustento teórico recayó sobre nosotros creemos que logramos abordar los tres ejercicios de manera correcta.

La parte de el análisis de los datos fue crucial , se basó en el escrito del proyecto. A pesar de desconocer la fuente real de los datos tratamos de plantear la función de hipótesis más adecuada para estos..

En cuanto al desarrollo de Machine Learning Tom Mitchell había escrito en su libro que: ?The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.?

Esto lo vimos reflejado en los tres ejemplos viendo como existía la necesidad de más interacciones para tener un mejor ajuste.