

Proyecto 3

Curso: Lenguajes de Programación

3 de octubre de 2018

Instituto Tecnológico de Costa Rica
Prof. Fabián Fallas Moya

Introducción

En este proyecto usted va a implementar regresión lineal y ver cómo funciona sobre datos.

Archivos para el proyecto (templates)

A través de este proyecto usted va a utilizar los scripts `proyecto4.m` y `proyecto4_multi.m`. Estos archivos van a inicializar y cargar los datasets y hacer las llamadas a las funciones necesarias para ejecutar todos los ejercicios. No es necesario modificar estos archivos, lo que sí es necesario es modificar las funciones que se encuentran en otros archivos, y seguir las instrucciones que se especifican en este documento.

Para la primera parte de este proyecto, usted trabajará con regresión lineal con una variable. La segunda parte de este proyecto, requiere el uso de regresión lineal con múltiples variables.

1. Calentamiento

En la primera parte vamos a practicar con poco con Octave. En la sección 1 del ejercicio, usted va a encontrarse con un template para insertarle código. Modifique esta sección para que retorne `A` como una matriz identidad 5×5 (no estamos recibiendo un parámetro).

Para no correr el resto del archivo `proyecto4.m` puede presionar los botones **ctrl** – **c** y la ejecución se detendrá.

2. Regresión lineal con una variable

En esta parte del ejercicio, usted implementará regresión lineal con una variable para predecir las ganancias de puestos de comida. Suponga que usted es el CEO de una franquicia de restaurantes y está considerando varias ciudades para abrir un nuevo restaurante. La cadena tiene varios restaurantes en ciudades y usted tiene información acerca de ganancias y poblaciones de dichas ciudades.

A usted le gustaría usar esta información para que le ayude a decidir cuál ciudad elegir para abrir el nuevo negocio.

El archivo `data1.txt` contiene un dataset para nuestro problema. La primera columna es la población de la ciudad y la segunda columna es la ganancia de un restaurante en esa ciudad. El archivo `proyecto4.m` carga el set de datos (comprendan el código).

2.1 Graficar la información

Antes de empezar, una herramienta muy útil es visualizar la información. Para esta ejercicio, se puede utilizar un gráfico de dispersión, ya que tenemos solamente dos propiedades (ganancia y población), aunque hay muchos problemas en donde graficar es más complejo (ya que algunos problemas tienen más de dos dimensiones). El resultado se muestra en la Figura 1.

Complete el archivo `graficar.m` con el siguiente código:

```
plot(x, y, 'rx', 'MarkerSize', 10);  
ylabel('Ganancia en $10,000s');  
xlabel('Población de la ciudad en 10,000s');
```

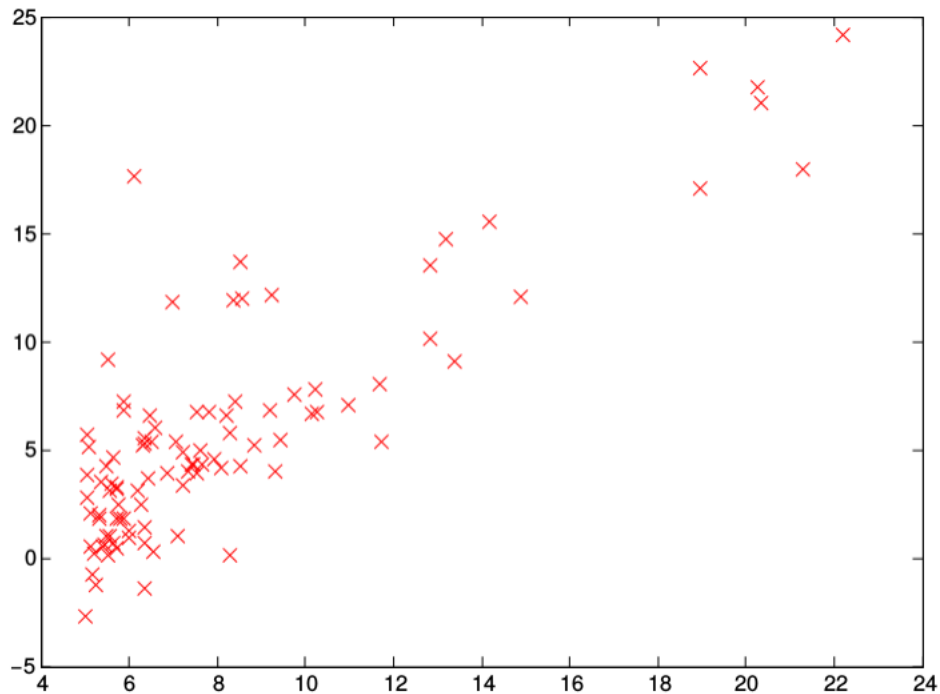


Figura 1: Gráfico de dispersión

2.2 Gradiente Descendente

En esta sección usted va a encontrar parámetros θ óptimos utilizando gradiente descendente.

2.2.1 Ecuaciones

Recuerde que el objetivo de regresión lineal es minimizar la función de costo $J(\theta)$. También recuerde que la función de hipótesis está dada por la fórmula: $h_{\theta}(x) = \theta^T x = \theta_0 + \theta_1 X$. Para poder obtener los parámetros θ , podemos utilizar gradiente descendente, en dónde lo que vamos a hacer es iterar utilizando la iteración:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Con cada iteración de gradiente descendente, los parámetros θ_j se acercan cada vez más a los valores óptimos para obtener el menor costo de la función $J(\theta)$.

2.2.2 Implementación

Para ir monitoreando si nuestro algoritmo está correcto, vamos a implementar una función que nos calcule el costo dado (con respecto a un par de parámetros θ). En esta sección usted va a calcular el costo de la función con respecto a parámetros θ . Complete el código en `costo.m`, para llevar a cabo lo anterior. Recuerde que las variables X y y no son valores escalares, sino matrices cuyas filas representan las «muestras» de nuestro training set.

Cuando los parámetros θ son ceros, el costo es de 32.07.

2.2.3 Gradiente descendente

Luego hay que completar el código en `gradienteDescendente.m`, en este lugar debe implementar la actualización de los parámetros θ (recuerde la actualización simultánea). Si dentro del ciclo usted va imprimiendo el costo de los parámetros, puede ser de ayuda ya que si usted ve que el costo se decrementa, eso quiere decir que la programación puede estar correcta. Al finalizar esta sección se imprimirá un gráfico similar al de la figura 2.

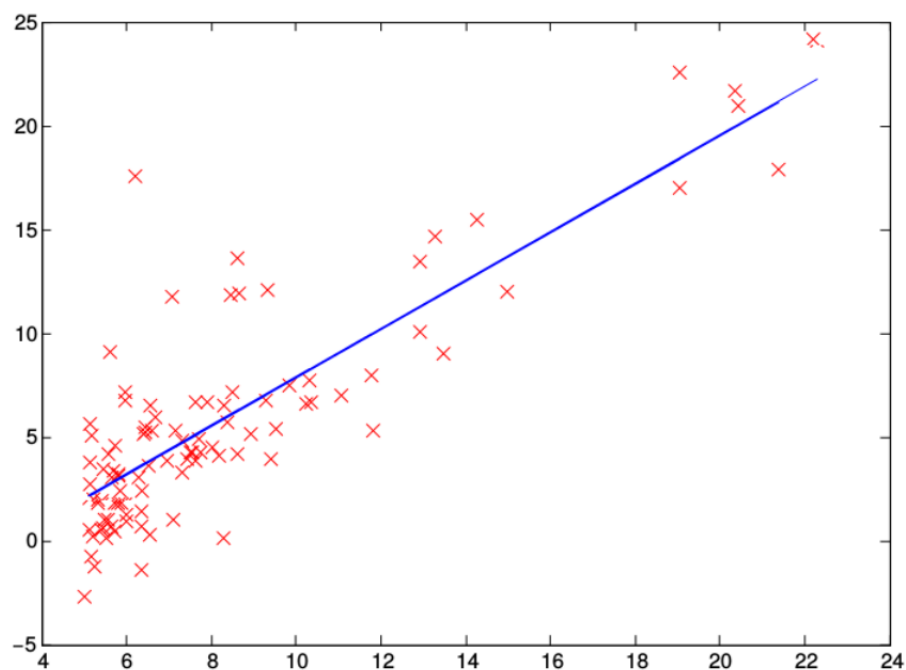


Figura 2: Función de hipótesis

2.2.4 Visualizando

En esta sección no debe modificar nada, solamente correr el ejercicio y comprender los gráficos que aparecen, en este caso el gráfico de contorno y otro gráfico de «superficie».

3. Regresión lineal con múltiples variables

Esta parte es similar a la anterior, solo que ahora hará uso de un dataset con más de una característica. Para esta se le proporciona el archivo `proyecto4_multi.m`, para que le sirva como template y poder llevar a cabo todas las tareas necesarias. Siga el orden de las secciones de ese archivo y ponga en práctica todo lo visto en clase.