Práctica 4. Regresión lineal y regresión logística

Competencia: el alumno será capaz de comprender e implementar modelos de regresión lineal y modelos de clasificación. Además, dominará el algoritmo de optimización gradiente descendente a través de su implementación en los modelos antes mencionados.

Descripción: implementar los ejercicios propuestos en el laboratorio haciendo uso eficiente de las herramientas.

Material:

- Computadora
- Anaconda
- Python 3.6
- Editor de texto
- Jupyter
- Numpy
- Matplotlib

La solución para cada uno de los ejercicios propuestos se debe hacer uso en manera de los posible de las bibliotecas numpy y matplotlib.

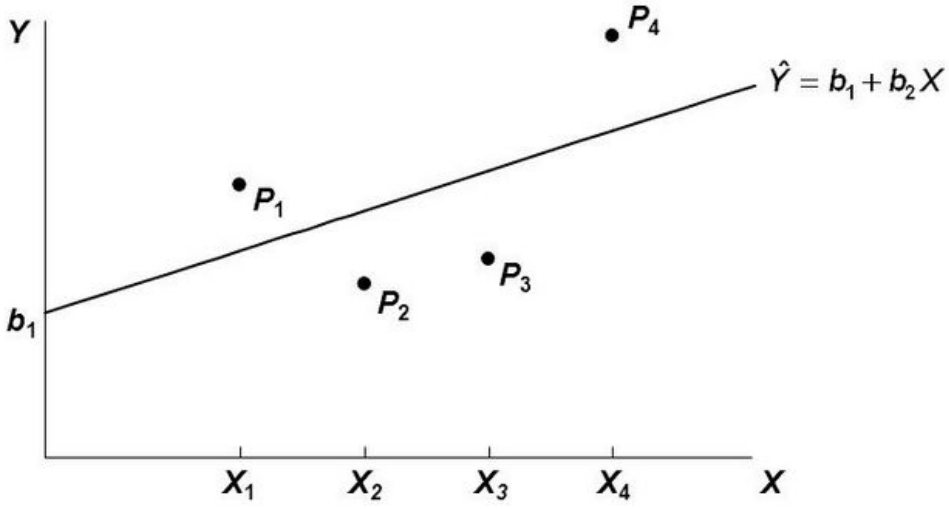
numpy es una biblioteca para python escrita en el lenguaje C, es una de las bibliotecas principales para el cómputo científico, provee de herramientas para manipulación de arreglos multidimensionales de una forma eficiente.

matplotlib es una biblioteca para python para graficar en 2D de alta calidad, ampliamente utilizada por los científicos para mostrar sus resultados en artículos científicos.

Se recomienda al estudiante investigar, a través de los sitios oficiales, las herramientas que disponen las siguientes bibliotecas: numpy, matplotlib, scipy, pandas.

Marco teórico

La regresión lineal consiste en el **ajuste** de una línea recta a un conjunto de observaciones definidas por puntos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.



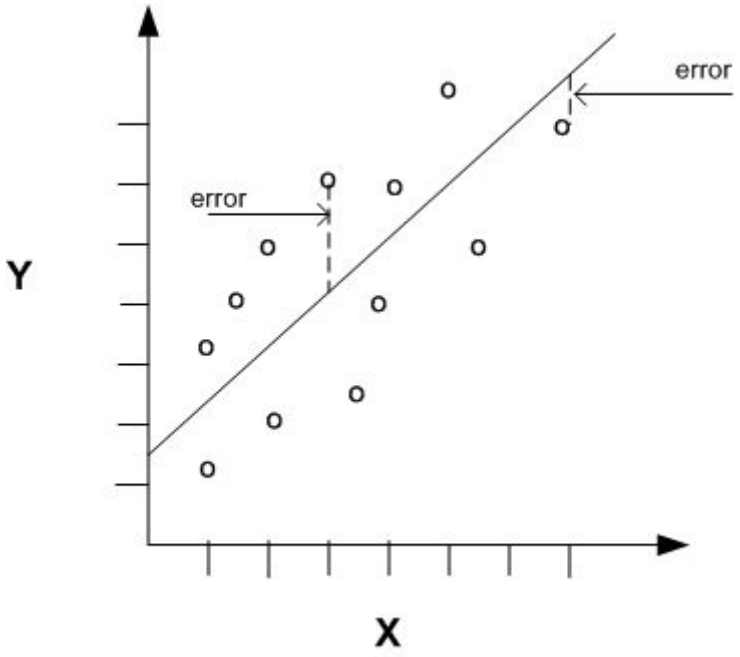
Expresión matemática para la línea recta es:

$$1)\ y = a_0 + a_1 x + e$$

donde a_0 y a_1 son coeficientes que representan la intersección con el eje y la pendiente, respectivamente, siendo e el error entre el modelo y las observaciones, despejando la ecuación (1) el error se puede representar como:

$$2)\ e = y - a_0 - a_1 x$$

El mejor ajuste de la línea recta al conjunto de datos consiste en minimizar el error e.



Ajuste de una línea recta por mínimos cuadrados

$$3)\ S_r = \sum_{i=1}^{n} e^2{}_i =$$

$$\sum_{i=1}^{n} (y_{i,\,medida} - y_{i,\,modelo})^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

NOTA:
El propósito de esta técnica es encontrar un línea recta que se ajuste mejor a nuestro conjunto de observaciones, los parámetros que se requieren encontrar para dicha línea recta son: a_0 y a_1.

Para encontrar los parámetros ideales siguiendo la técnica de mínimos cuadrados se debe derivar parcialmente la ecuación del error (3) respecto a cada uno de sus coeficientes.

Derivada parcial de la ecuación (3) respecto al parámetro a_0

$$4)\ \frac{\eth S_r}{\eth a_0} = -2 \sum (y_i - a_0 - a_1 x_i)$$

Derivada parcial de la ecuación (3) respecto al parámetro a_1

$$5)\ \frac{\eth S_r}{\eth a_1} = -2\ \sum[(y_i - a_0 - a_1 x_i)\, x_i]$$

Para determinar la expresión matemática que minimice el error, ambas ecuaciones 4 y 5 se igualan a 0. Se debe tener en mente que las sumatorias van desde *i*=1 hasta *n*, donde *n* es el número de observaciones.

Ecuación 4 igualada a 0.

$$6)\ 0 = \sum y_i - \sum a_0 - \sum a_1 x_i$$

Ecuación 5 igualada a 0.

$$7)\ 0 = \sum y_i x_i - \sum a_0 x_i - \sum a_1 x_i^2$$

El término $\sum a_0$ nótese que se puede expresar como $n * a_0$, ahora expresamos las ecuaciones como un conjunto de dos ecuaciones lineales simultáneas con dos incógnitas, ecuaciones 8 y 9.

$$8)\ na_0 - (\sum x_i)a_1 = \sum y_i$$

$$9)\ (\sum x_i)a_0 - (\sum x_i^2)a_1 = \sum y_i x_i$$

Resolviendo el sistema de ecuaciones obtenemos las ecuaciones para determinar los parámetros $a_0\, y\, a_1$, ecuaciones 10 y 11.

$$10)\ a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$11)\ a_0 = \bar{y} - a_1 \bar{x}$$

donde $\bar{y}$ y $\bar{x}$ son las medias de y y x respectivamente.

Estimación del error del modelo obtenido

Suma de los errores al cuadrado

$$sse = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

donde $\hat{y}_i$ es la salida del modelo.

Error cuadrático medio

$$mse = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{sse}{n}$$

Error de la raíz cuadrada de la media

$$rmse = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{mse} = \sqrt{\frac{sse}{n}}$$

Desarrollo de la práctica

Leer detenidamente la presentación “Unidad 3. Regresión Lineal y Regresión Logística” para el desarrollo de los siguientes ejercicios:

Ejercicio 1.
Utilizar las bibliotecas de numpy y matplotlib para resolver el siguiente problema propuesto. Implementar el método de ajuste de parámetros, mínimos cuadrados, para encontrar los parámetros óptimos de una función lineal. Se deberá imprimir los puntos a ajustar (tabla 1) y la salida del modelo (una línea recta).

Ejercicio 2.
Utilizar las bibliotecas de numpy y matplotlib para resolver el siguiente problema propuesto. Implementar el método de ajuste de parámetros, gradiente descendente, para encontrar los parámetros pseudo-óptimos de una función lineal. Se deberá imprimir los puntos a ajustar (tabla 1) y la salida del modelo (una línea recta). Comparar el error generado respecto al ajuste por mínimos cuadrados.

Tabla 1. Datos de ajuste

Xi	Yi
1	0.50
2	2.50
3	2.00
4	4.00
5	3.50
6	6.00
7	5.50

Ejercicio 3.
Utilizar las bibliotecas de numpy y matplotlib para resolver el siguiente problema propuesto. Implementar el modelo de regresión logística para clasificar los datos de la **tabla 2**. Emplear el método de ajuste de parámetros, gradiente descendente, para encontrar los parámetros pseudo-óptimos para este modelo. Se deberá imprimir los puntos a ajustar (tabla 2) y la salida del modelo (una sigmoide).

Tabla 2. Datos de ajuste

Aprobó	Horas
0	0.50
0	0.75
0	1.0
0	1.25
0	1.5
0	1.75
1	1.75
0	2.0
1	2.25
0	2.5
1	2.75
0	3.0
1	3.25
0	3.50
1	4.0
1	4.25
1	4.5
1	4.75
1	5.0

1	5.5

Contenido del reporte

El reporte de la práctica debe consistir en los siguientes apartados para cada ejercicio:

- Descripción del ejercicio
- Análisis matemático (si lo contiene; ej. derivadas)
- Funciones utilizadas (ej. **numpy.dot**)
- Tabla de valores obtenidos
- Gráficas
- Código (con formato)