Aprendizaje Automático: Regresión logística y polinomial

M. Sc. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Programa de Ciencias de los Datos,
PAttern Recongition and MAchine Learning Group (PARMA-Group)

16 de julio de 2019

Fecha de entrega: Domingo 28 de Julio.

Entrega: Un archivo .zip con el código fuente LaTeX o Lyx, el pdf, y el script en jupyter, debidamente documentado, con una función definida por ejercicio. A través del TEC-digital.

Modo de trabajo: Grupos de 2 personas.

En el presente trabajo práctico se introducirá el problema de la regresión con modelos polinomiales y logística.

1. (40 puntos) Implementación de clasificadores lineales de dos clases

1. Para facilitar la implementación del presente laboratorio, se provee el archivo *PerceptronPytorch.*py, el cual implementa la generación de datos aleatorios **linealmente separables**.

2. **Algoritmo de Regresión logística**:

   a) **(10 puntos)** Implemente el algoritmo de regresión logística rescindiendo al máximo de estructuras de tipo *for,* usando entonces operaciones matriciales.

   b) **(5 puntos)** Cuál de las funciones matemáticas vistas en clase puede utilizarse a la salida del regresor logístico para realizar una clasificación binaria? Explique cómo modificar la regresión logística entonces para realizar una clasificación binaria.

   c) **(25 puntos)** Realice 2 pruebas con distintas distancias de separación entre las muestras de las clases, con una prueba linealmente separable, y otra no, y documente el número (en una tabla) de muestras mal clasificadas y la cantidad de iteraciones para converger. Defina

el conjunto de muestras de entrenamiento como el 70 % de las muestras aleatoriamente seleccionadas, y el resto utilicelas como muestras de prueba.

1) Reporte los resultados promedio para 10 corridas, con promedios y desviaciones estándar.

2. (60 puntos) Experimentos, análisis y el enfoque de mínimos cuadrados regularizado

1. **(10 puntos)** Genere una función como la mostrada en la ecuación

$$t = \sin(2\pi x) + \epsilon(x), \tag{1}$$

en la que el seno se contamine con ruido Gaussiano, con relación señal a ruido de 4 dB (puede usar la función *awgn* para contaminar con ruido Gaussiano tal señal unidimensional).

   a) Utilice dos conjuntos de datos con $N = 10$ y $N = 100$ para todos los puntos posteriores.

   b) Grafique la salida de la función t en el intervalo $[0, 1]$, usando la función *scatter* de *pyhon*.

2. **(30 puntos)** Programe la función $getOptimumW(\vec{x}, \vec{t}, M)$, la cual obtenga el arreglo óptimo de pesos $\vec{w}_{\text{opt}}$ usando el enfoque de mínimos cuadrados, donde $\mathcal{D} = \{\vec{x}, \vec{t}\}$ es el conjunto de datos de entrenamiento obtenidos en el punto anterior.

   a) Experimente con al menos cuatro valores distintos de M, grafique el modelo resultante y la función sin ruido, y explique el fenómeno de sobreajuste.

   b) Presente una tabla por cada valor de M en la que se evalúe el error RMS, tanto para el conjunto de datos de entrenamiento (70 %), como el de validación (30 %) como soporte para la explicación solicitada anteriormente

$$E\left(\vec{w}_{\text{opt}}\right)_{\text{RMS}} = \sqrt{2E\left(\vec{w}_{\text{opt}}\right)/N}. \tag{2}$$

3. **(20 puntos)** Para atenuar el problema de sobre-ajuste, el enfoque de mínimos cuadrados regularizados propone agregar un término que «castigue» los vectores de pesos $\vec{w}$ con dimensionalidad alta, sumando la norma de $\vec{w}$ y pesada por el parámetro de regularización λ:

$$E(\vec{w}) = \frac{1}{2}\sum_{n=0}^{N}\left\{y\left(x_n, \vec{w}\right) - t_n\right\}^2 + \frac{\lambda}{2}\left\|\vec{w}\right\|^2. \tag{3}$$

Implemente el enfoque de mínimos cuadrados regularizado por el parámetro λ, y pruebe al menos tres valores distintos para tal parámetro (con uno de ellos $\lambda = 0$), con dos conjuntos de datos de $N = 10$ y $N = 100$.

a) Grafique el modelo resultante para cada variante, junto con la función sin ruido.

b) Evalúe la función de error RMS para cada caso, y presente una tabla con los tres valores de λ y los dos valores de N escogidos.

c) Explique el comportamiento observado, y relaciones la manipulación de λ respecto a M en el apartado anterior, y la cantidad de datos N. ¿Qué relación encuentra entre λ y N?