Trabajo práctico 2: Regresión Logística

Ph. D. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Ingeniería en Computación
PAttern Recongition and MAchine Learning Group (PARMA-Group)

13 de mayo de 2024

Fecha de entrega: Lunes 3 de Junio
Entrega: Un archivo .zip con el código fuente LaTeX o Lyx, el pdf, y un notebook Jupyter, debidamente documentado, con una función definida por ejercicio. A través del TEC-digital.
Modo de trabajo: Grupos de 3 personas.

Resumen

En el presente trabajo práctico se introduce la implementación de la selección de características y la regresión logistica para fines de clasificación.

1. Regresión Logística con datos sintéticos (60 puntos)

1. **(40 puntos)** Algoritmo de Regresión logística: Implemente el algoritmo de regresión logística rescindiendo al máximo de estructuras de tipo *for*, usando entonces operaciones matriciales.

2. Para cada clasificador:

   a) **(20 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 (en una tabla) el error medio absoluto (MAE), el cual se define como sigue:

$$e_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}\left|t_i - \widetilde{t}_i\right|.$$

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

2) Documente la cantidad de iteraciones usadas, y grafique la función del error durante el entrenamiento, para una corrida de ambos casos (datos linealmente separables y no linealmente separables). Explique los resultados.
3) Grafique con un diagrama de dispersión los datos generados, junto con la superficie de decisión construida por el modelo en cada caso (definida según los pesos). **Explique los pasos para graficar la superficie de decisión**.

b) **(40 puntos)**Reporte los resultados promedio para 10 corridas, con los datos linealmente separables y no linealmente separables generados de forma aleatoria en cada una, en clasificar tales muestras de prueba.

1) Reporte el promedio del MAE y su desviación estándar para las 10 corridas, en una tabla. Comente los resultados.

2. Regresión Logística para clasificación de ataques (40 puntos)

Basado en el articulo "*UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection systems*" utilice su implementación del algoritmo de regresión logística para detectar los ataques, usando el dataset publicado en tal articulo:

1. Implemente usando las funciones básicas de pytorch las métricas de aciertos, falsos positivos, falsos negativos, verdaderos positivos, verdaderos negativos, precision, *recall* y el F1-score.

   a) Diseñe al menos dos pruebas unitarias donde compare los resultados con las métricas implementadas en el paquete *sklearn.metrics.*

2. Implemente el preprocesamiento de los datos de la siguiente forma: calcule la media y desviación de estándar de la partición de datos de entrenamiento, y extraiga la media de los datos para luego dividir los mismos por la desviación estándar.

3. Realice 30 particiones distintas de datos de entrenamiento y prueba, usando la función *train_test_split* de *sklearn*. Muestre los resultados al entrenar el modelo de regresión logística construido en los puntos anteriores, en una tabla, junto con la media y desviación estándar.

   a) Ejecute el entrenamiento usando el preprocesamiento del punto anterior y sin usarlo.

   b) Compare y comente los resultados.