Inteligencia artificial avanzada para la ciencia de datos I

TC3006C

Dr. Esteban Castillo Juarez
TEC de Monterrey, Campus Santa Fe

esteban.castillojz@tec.mx

Tecnológico de Monterrey

Agenda

- Objetivo
- Instrucciones
- Conjunto de datos sobre el cáncer de mama
- Formato de la actividad

Tecnológico de Monterrey

Objetivo

- Reforzar la teoría de regresión logística aplicando el algoritmo de clasificación en un conjunto de datos diferente.

- Practicar la teoría detrás de la regresión logística en el contexto del lenguaje de programación Python con plottly express y scikit-learn.

- Aprender la teoría detras de una matriz de confusión y las métricas de precisión y recuerdo.

Tecnológico de Monterrey

Instrucciones

Para la implementación manual y con Scikit-learn vista en clase hacer lo siguiente:

1. Utilice el conjunto de datos de “cáncer de mama” para entrenar una nueva versión del clasificador de regresión logística.

2. Adapte el conjunto de datos al formato correcto que necesita el algoritmo. Sugerencia: cambie la etiqueta de clasificación (benigno y maligno) a 0.0 y 1.0.

3. Proponga una nueva función de activación como ReLU, Tanh, entre otras.

Instrucciones

Para la implementación manual y con Scikit-learn vista en clase hacer lo siguiente:

4. Cambien el numero de iteraciones de 1 a 150 así como cambien Alpha en el rango de 0.01 a 0.30 (en el código manual, alpha es lo siguiente alpha=4/(1.0+x+y)+**0.01**). Esto, ¿se puede en la implementación de scikit-learn? ¿Cómo se haría con los hiperparametros de la herramienta?

5. Compare la precisión y el recuerdo de cada implementación hecha y vea cual es mejor así como genere la visualización de una matriz de confusión promedio.

Instrucciones

Para la implementación manual y con Scikit-learn vista en clase hacer lo siguiente:

6. Proponga cuatro visualizaciones en plotly express que le ayuden a entender la naturaleza del modelo de regresión logística.

6. Describa los hallazgos y proporcione una conclusión sobre el proceso de experimentación.

Tecnológico de Monterrey

Conjunto de datos sobre el cáncer de mama

- El conjunto de datos sobre cáncer de mama se obtuvo del repositorio de la UCI. Puede consultarlo para obtener información más detallada.

- El conjunto de datos proviene originalmente del Hospital Universitario de Wisconsin, del Dr. William H. Wolberg.

- Las muestras se almacenaron periódicamente a medida que el Dr. Wolberg informaba sobre sus casos clínicos. Por lo tanto, la base de datos refleja una agrupación cronológica de datos.

Tecnológico de Monterrey

Conjunto de datos sobre el cáncer de mama

```
Group 1: 367 instances (January 1989)
Group 2:  70 instances (October 1989)
Group 3:  31 instances (February 1990)
Group 4:  17 instances (April 1990)
Group 5:  48 instances (August 1990)
Group 6:  49 instances (Updated January 1991)
Group 7:  31 instances (June 1991)
Group 8:  86 instances (November 1991)
-----------------------------------------
Total:   699 points (as of the donated datbase on 15 July 1992)
```

Tecnológico de Monterrey

Conjunto de datos sobre el cáncer de mama

- Número total de instancias: 699 (al 15 de julio de 1992).
- Número de instancias de entrenamiento: 599.
- Número de instancias de prueba: 100.
- Número de atributos: 9 más el atributo de clase (última columna).
- Todos los elementos están separados por comas.

Tecnológico de Monterrey

Conjunto de datos sobre el cáncer de mama

#	Attribute	Domain
1.	Clump Thickness	1.0 - 10.0
2.	Uniformity of Cell Size	1.0 - 10.0
3.	Uniformity of Cell Shape	1.0 - 10.0
4.	Marginal Adhesion	1.0 - 10.0
5.	Single Epithelial Cell Size	1.0 - 10.0
6.	Bare Nuclei	1.0 - 10.0
7.	Bland Chromatin	1.0 - 10.0
8.	Normal Nucleoli	1.0 - 10.0
9.	Mitoses	1.0 - 10.0
10.	Class:	(benign,malignant)

Class distribution:

Benign: 458 (65.5%)
Malignant: 241 (34.5%)

Tecnológico de Monterrey

Formato de la actividad

- Todas las actividades son individuales.
- La actividad se entregará el día de la siguiente sesión de clases con el profesor.
- Todas las tareas deben estar documentadas en detalle. Las tareas deben entregarse o enviarse como un archivo comprimido (nombre: matricula) con la siguiente información:
  - Archivo de Python, notebook de Jupyter o Google Colab.
  - Documento con explicación del problema e imágenes de su ejecución. Pueden simplemente mostrar un notebook (si lo ocuparon) si toda la documentación está en celdas de markdown, debidamente explicada.

Puede usar Github con markdown para documentar el proceso y solo enviar la URL.

Formato de la actividad

- La actividad deben contener lo siguiente:
  - Portada con información del estudiante (5%).
  - Sección de introducción que explica explícitamente el problema abordado (10%).
  - Sección experimental que explica el enfoque utilizado para resolver el problema (30%).
  - Sección de resultados que recoge los hallazgos hechos(40%).
  - Sección de conclusiones que recoge lo aprendido en la actividad (15%).

Tecnológico de Monterrey

¡Gracias!

Tecnológico de Monterrey