Semestral AD2021

Aprendizaje automático
Práctica 2 “Regresión logística”

La práctica se deberá realizar en equipos de 2 personas.

Datasets

1. Dataset 1: Credit Card Default Data (`Default.txt`). Este conjunto de datos contiene 10,000 instancias donde el objeto es predecir cuáles clientes van a incumplir con la deuda de su tarjeta de crédito. Este dataset cuenta con el siguiente formato.
    a. `ID:` número de la entrada (descartar).
    b. `default:` un factor con niveles `No` y `Yes` indicando si el cliente va a incumplir su deuda. Este es la variable de salida del dataset.
    c. `student:` un factor con niveles `No` y `Yes` indicando si el cliente es un estudiante. Actúa como atributo.
    d. `balance:` el saldo promedio que le queda al cliente en su tarjeta de crédito después de realizar su pago mensuale. Actúa como atributo.
    e. `income:` ingreso del cliente. Actúa como atributo.
2. Dataset 2: Identificación de género (`genero.txt`). Este conjunto contiene 10,000 instancias donde el objeto es predecir el género de una persona, es decir, `Male` o `Female`. Este dataset cuenta con el siguiente formato.
    a. `Gender:` género de la persona con dos niveles: `Male` o `Female`.
    b. `Height:` altura de la persona.
    c. `Weight:` peso de la persona.

A continuación, se presentan los procedimientos para la regresión logística usando Scikit-Learn y empleando el método de gradiente descendente. El procedimiento debe ser realizado para los dos datasets. Los datasets deberán ser partidos de forma aleatoria en una proporción 80%-20% para generar el training set y test set, respectivamente. En todos los casos se hará uso de una validación simple.

Procedimiento para regresión logística con Scikit-Learn

1. Investigar el uso de la regresión logística en `Scikit-Learn`, es decir, la función `LogisticRegression()`.
2. Entrenar el modelo.
3. Calcular la precisión de los modelos (obtener la tasa de precisión) y la matriz de confusión con `Scikit-Learn`.

4. Sólo para el caso del dataset de identificación de género se debe realizar lo siguiente.
    a. Con base en el test set que se utilizó para generar los modelos, graficar las instancias que lo constituyen, es decir graficar `Height vs Weight`. Cada punto debe ser identificado de acuerdo con la clase que tiene asociada. Color rojo para `Female` y azul para `Male`.
    b. Realizar este tipo de graficación ahora con base en lo que predijo el modelo.

Procedimiento para regresión logística con Gradiente Descendente

1. Programar el gradiente descendente, empleando el gradiente visto en clase: $\nabla J(\vec{\beta}_j) = X^T(\vec{\mu} - \vec{y})$, donde el i-ésimo elemento del vector $\vec{\mu}$, es decir, $\mu_i = 1/\left(1 + e^{-\vec{\beta}_j^T \cdot \vec{x}_i}\right)$
2. Entrenar el modelo por cada dataset.
3. Calcular la precisión de los modelos (obtener la tasa de precisión) y la matriz de confusión con `Scikit-Learn`.
4. Sólo para el caso del dataset de identificación de género se debe realizar lo siguiente.
    a. Con base en el test set que se utilizó para generar los modelos, graficar las instancias que lo constituyen, es decir graficar `Height vs Weight`. Cada punto debe ser identificado de acuerdo con la clase que tiene asociada. Color rojo para `Female` y azul para `Male`.
    b. Realizar este tipo de graficación ahora con base en lo que predijo el modelo.
5. Reportar el vector $\vec{\beta}$ obtenido por el método de gradiente descendente, así como también informar cuál criterio de paro se usó y la tasa de aprendizaje α empleada.

Reporte de la práctica

Emplear el formato de práctica dado por el profesor y seguir las instrucciones mostradas. El archivo que se subirá a *Canvas* deberá estar estrictamente en formato PDF y deberá ser nombrado como `report.pdf`.

Usar el lenguaje `Python` para desarrollar la práctica. **Únicamente será aceptado este lenguaje para la generación de los programas**. Además, es forzoso el uso de `Scikit-learn`. Entregar el programa con extensión `.py` debidamente comentado. Entregar un archivo `README.txt` donde se exponga cómo ejecutar el programa (indicar los parámetros en caso de necesitarlos) y un ejemplo para cómo ejecutar el programa y producir así los resultados reportados.

Entrega global

Tanto el reporte y el programa deberán ser empaquetados en un archivo `.ZIP` y nombrarlo: `practice2.zip`. **Cualquier falta a las instrucciones pedidas implicará la anulación de la práctica para todos los integrantes del equipo.**