Universidad De Granada

Apredizaje Automático

Proyecto Final

Pedro Bonilla Nadal
Antonio Martín Ruiz

11 de junio de 2020

Índice

1. Compensión del problema a resolver

Este dataset contiene datos de la oficina del censo[1], relativos al censos de 1995, obtenidos de la repositorio UCI de bases de datos [2]. Un total de 48842 personas censadas han sido encuestadas para este censo. De estas personas tenemos una seríe de variables con información de carácter socio económico. En particular:

- tenemos 6 variables de tipo numérico y entero, con valores en rangos distintos.
- 8 variables de tipo categórico.
- Una variable de clase que toma como valores < 50K y > 50k.
- Es un problema de clasificación, ya que no tenemos información para hacer una regresión sobre la ganancia.

El código utilizado para la resolución de la práctica se encuentra en el archivo `main.py`. Del mismo modo, los datos limpiados se encuentran en el archivo `adults.data` y su información relativa procesada en el archivo `adult.names`.

2. Disvisión y Codificación de los conjuntos

A la hora de obetener los datos obtenemos tres archivos del repositorio: `adult.data`, `adult.test` y `adult.names`. Estos datos fueron procesados en el año 1996. Nosotros realizamos una operación de formato de los datos para ajustarlo a convenciones más recientes.

- Cambiamos la variable de clase por valores categóricos 0 para < 50K y 1 para > 50K
- Eliminamos el espaciado después de la coma, para facilitar la lectura por parte de librerías actuales.
- Añadimos la información el archivo `adult.test` al archivo `adult.data` con el objetivo de tener la información procesada de manera conjunta.
- Añadimos el flag `@attribute` a la lista de atributos provista en `adult.names`.

Una vez leidos los datos, haremos una division train/test de 80 %/20 %. Elegimos esta proporción aprovechando que dada la gran cantidad de datos de los que poseemos. Obtenemos por lo tanto un conjunto de train con 39074 instancias, así como un conjuto de test con 9768 instancias.

Además, a la hora de validar la selección de hiperparámetros utilizaremos la técnica CrossValidation, para la cual diviremos el conjunto de training en 5 subconjuntos.

3. Preprocesado

3.1. Valoración de las variables de interés

3.2. Normalización de las variables

3.3. Clases de funciones?

4. Función de pérdida usada

5. Técnica de Ajuste para el Modelo Lineal

Selección de las técnica (parámetrica) y valoración de la idoneidad de la misma frente a otras alternativas

6. Hiperparámetros y selección del modelo

Aplicación de la técnica especificando claramente que algoritmos se usan en la estimación de los parámetros, los hiperparámetros y el error de generalización

7. Error de Generalización

8. Argumentar sobre la idoneidad de la función regularización usada

9. Conclusiónes

Valoración de los resultados y justificación (gráficas, métricas de error, análisis de residuos, etc)
que se ha obtenido la mejor de las posibles soluciones con la técnica elegida y la muestra dada. Argumentar en términos de los errores de ajuste y generalización,

Referencias

[1] Página oficial de la Oficina del censo: http://www.census.gov/ftp/pub/DES/www/welcome.html.

[2] Página del Centro para el Apredizaje Automático y Sistemas Inteligentes:http://archive.ics.uci.edu/ml/datasets/Adult.