

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	2
2. División y codificación de los conjuntos	2
3. Preprocesado	2
4. Función de pérdida usada	3
5. Técnica de Ajuste para el Modelo Lineal	3
6. Hiperparámetros y selección del modelo	3
7. Error de Generalización	4
8. Argumentar sobre la idoneidad de la función regularización usada	4
9. Conclusiones	4

## 1. Compensión del problema a resolver

Este dataset contiene datos de la oficina del censo[1] relativos a censos de 1995, obtenidos del repositorio UCI de bases de datos [2]. Un total de 48842 personas han sido encuestadas para este censo. De estas personas tenemos una serie de variables con información de carácter socioeconómico. 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. División y codificación de los conjuntos

A la hora de obtener los datos en el repositorio se encuentran tres archivos: `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 ajustarlos 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 del 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 leídos los datos haremos una división 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 conjunto de test con 9768 instancias.

Además, a la hora de validar la selección de hiperparámetros utilizaremos la técnica cross validation, para la cual dividiremos el conjunto de training en 5 subconjuntos. Se utilizará para esta la función `cross_validate`[4].

## 3. Preprocesado

Mostraremos a continuación las siguientes técnicas de preprocesado realizadas. Justificaremos individualmente su uso.

- Normalización de las variables: realizaremos una normalización de los datos, para evitar que su escala afecte a la relevancia de estos. Para ello usaremos la función `StandardScaler`[3] provista en la librería `sklearn`.
- Aumento de dimensionalidad: debido a la baja cantidad de variables, en contraposición con la alta cantidad de ejemplos podemos considerar aumentos de la dimensionalidad de los datos sin miedo a que se genere un gran sobreajuste.

- Dummy Variables: una variable dummy es aquella que toma sólo el valor 0 o 1 para indicar la ausencia o la presencia de algún efecto categórico que se pueda esperar que cambie el resultado. Esta técnica ayudará a codificar la entrada para algunos algoritmos, así como poder hacer un estudio de cada categoría como una variable separada a la hora de hacer una valoración del interés de cada variable.

Incluir información de como se ha realizado.

- Polynomial Features: con especial interés en el caso lineal, el uso de variaciones polinómicas de los datos puede ser útil para permitir aproximaciones a clases de funciones nuevas. Probaremos variación cuadrática sobre las variables numéricas cuando realizemos un ajuste lineal. Para Random Forest no lo consideramos necesario por la ya conocida complejidad de los árboles, y para SVM utilizaremos variaciones de kernell.

Incluir información de como se ha realizado

- Valoración de las variables de interés:
- Datos incompletos y valores perdidos: en relación a los valores perdidos encontramos en tres variables de tipo categórico.

Variable	Valores distintos	Valores Perdidos
<code>workclass</code>	9	2799
<code>occupation</code>	15	2809
<code>native-country</code>	42	857

Tabla 1: Representación de valores perdidos.

Los sustituimos haciendo esto.

- Datos inconsistentes: no encontramos clases inconsistentes.

## 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. Conclusiones**

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 Aprendizaje Automático y Sistemas Inteligentes: <http://archive.ics.uci.edu/ml/datasets/Adult>.
- [3] Función `StandardScaler` de la librería `sklearn`: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
- [4] Función `cross_validate` de la librería `sklearn`: [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.cross\\_validate.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html)
- [5] Artículo donde explica el concepto de dummy variable: [https://en.wikipedia.org/wiki/Dummy\\_variable\\_\(statistics\)](https://en.wikipedia.org/wiki/Dummy_variable_(statistics))