Práctica ML 101

Esta práctica, diseñada por Felipe Alonso Atienza

Objetivo

El objetivo de la práctica es simple: abordar un problema de machine learning *realista* siguiendo la metodología y buenas prácticas explicadas durante las clases teóricas.

Por tanto, en estas instrucciones no se especifican los pasos exactos que el alumno tiene que llevar a cabo para realizar esta tarea con éxito. Es parte del trabajo aplicar las técnicas de procesamiento/transformación de variables que mejor se acondicionen al problema, identificar los modelos adecuados que proporcionen buenas prestaciones así como las variables potencialmente más relevantes, y elegir la métrica adecuada para contrastar los distintos modelos.

Las posibilidades son amplias, así que es recomendable abordar una aproximación incremental, esto es, comenzar por soluciones sencillas para progresivamente aumentar la complejidad de las técnicas utilizadas.

Datos

Se tendrán que abordar dos tareas de aprendizaje automático: un problema de clasificación y un problema de regresión. En ambos casos, se proporciona un **conjunto de entrenamiento**, y un **conjunto de test sin las etiquetas de la variable objetivo**.

1. Un problema de clasificación

Archivos: `census_train.csv, census_test.csv`

Este conjunto de datos es una versión modificada del utilizado en el artículo *"Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid"*, escrito por Ron Kohavi. Los datos originales se pueden encontrar en el *UC Irvine Machine Learning Repository*.

El objetivo es predecir si los ingresos de una persona superan o no los 50.000$ (variable *income*). Para ello, se tienen 13 características:

- ***age***: Edad
- ***workclass***: tipo de ocupación (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)
- ***education_level***: Nivel educativo (Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool)

- ***education-num***: Número de años de educación completados.
- ***marital-status***: estado civil (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)
- ***occupation***: ocupación (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces)
- ***relationship***: familia (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
- ***race***: raza (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
- ***sex***: Género (Female, Male)
- ***capital-gain***: Ganancia de capitales
- ***capital-loss***: Pérdida de capitales
- ***hours-per-week***: Promedio de horas trabajadas por semana
- ***native-country***: País de origen (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands)

2. Un problema de regresión

Archivos: `diamonds_train.csv, diamonds_test.csv`

El objetivo es estimar el precio (*price*) de un diamante a partir de un conjunto de propiedades físicas del mismo:

- ***carat***: peso del diamante
- ***cut***: calidad del corte: Fair, Good, Very good, Premium, Ideal.
- ***color***: desde color J (peor color) a D (mejor color)
- ***clarity***: medida de la claridad del diamante (I1 (peor calidad), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (mejor claridad))
- ***x***: longitud en mm
- ***y***: anchura en mm.
- **z**: profundidad en mm
- ***depth***: porcentaje de profundidad total 2 * z / (x + y)
- ***table***: ancho de la punta del diamante relativa al punto más ancho.

Formato de entrega

El alumno abordará los dos problemas en ficheros **Jupyter Notebook** (extensión .ipynb) independientes, que serán almacenados en la carpeta drive correspondiente y cuyo enlace se incluirá en el fichero de revisión de prácticas.

Este archivo no debe de contener únicamente código, sino que **han de justificarse convenientemente todos los pasos dados y las decisiones tomadas**. Se valorará positivamente por tanto, la comunicación de resultados (vía gráficas y explicaciones escritas), así como la interpretación de los resultados obtenidos.

IMPORTANTE: Adicionalmente a la resolución, el alumno deberá entregar dos archivos csv (separados por ","), con el resultado del algoritmo para el conjunto de test proporcionado. Este fichero CSV tendrá dos columnas:

- **id: representa el identificador asociado del conjunto de test proporcionado.**
- **target: la variable de salida estimada con el algoritmo desarrollado.**

Fecha de entrega

La fecha de entrega seguirá las condiciones fijadas en el Bootcamp.