

Descripción de la práctica y criterios de corrección

Fundamentos de Aprendizaje Automático

Esta segunda parte de la materia realiza un enfoque de proyectos de aprendizaje. Es decir, durante la misma se deberá escoger un proyecto de interés por parte de los grupos de estudiantes el cual desarrollaran durante la segunda parte de las prácticas del curso. El objetivo, por tanto, de esta segunda práctica es doble. En primer lugar, familiarizarse con los distintos problemas en los que se pueden aplicar técnicas de AA en un enfoque más “clásico”, es decir, sin entrar en terreno de proceso de imágenes, señales o lenguaje natural. El segundo objetivo es adquirir experiencia en la aplicación de las técnicas de AA en la resolución de un problema del mundo real, utilizando para ello las funciones desarrolladas en la práctica 1. Dado que estas funciones han sido desarrolladas para resolver problemas de clasificación, esta segunda práctica se centrará en este tipo de problema.

Para alcanzar estos objetivos, es necesario partir de un conjunto de datos, para lo cual pueden descargarse distintas bases de datos de internet. El sitio de internet más conocido que contiene datos para la aplicación de técnicas de *machine learning* es el repositorio de la Universidad de California, en la siguiente dirección:

<https://archive.ics.uci.edu/>

En esta página se puede encontrar una gran cantidad de bases de datos con las descripciones de distintos problemas a resolver. Por tanto, un primer trabajo será explorar alguno de estos sitios, si bien se recomienda principalmente el repositorio UCI dada su simplicidad, y analizar distintas bases de datos para escoger una de ellas, que se utilizará en esta segunda práctica. Es importante tener en cuenta que este problema debe de ser de **clasificación**, y se debe consultar con el profesor si el problema escogido es adecuado, para descartar aquellos de complejidad excesiva o que presenten algún otro problema.

En este sentido, es mejor descartar aquellos problemas en los que las bases de datos están formadas por imágenes o señales, puesto que sería necesario procesarlas para extraer las características que se utilicen de entrada al modelo. Problemas válidos sí podrían ser aquellos que partan de imágenes o señales, pero que la base de datos a descargar no contenga estas imágenes o señales sino características extraídas de las mismas.

Otros problemas que se podrían descartar pueden ser aquellos que tengan un número muy bajo de patrones (por ejemplo, inferior a 100), o excesivamente alto (por ejemplo, superior a 10000 patrones).

Además, descartamos aquellos problemas en los que falten ciertos valores (*missing values*), y los que no están relacionados con los tipos de problemas que vamos a intentar resolver (clasificación).

Al hacer clic en uno de los problemas, aparece en primer lugar una descripción en la que figura, entre otras cosas, el número de patrones (# *Instances*), número de atributos (# *Features*), el tipo de problema que se quiere resolver (*Associated Tasks*), la naturaleza de las características que tiene el problema (*Data Set Characteristics*), y si faltan valores o no (*Missing Values*). Más abajo aparece una descripción más detallada de la base de datos con información sobre cada entrada y salida. Una vez verificado que este problema es válido, se puede descargar la base de datos. La siguiente imagen muestra la descripción de una base de datos con 150 patrones, con 4 atributos (4 entradas), sin ningún valor que falte, para un problema de clasificación:



Iris

Donated on 6/30/1988

A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluating classification methods.

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Biology	Classification
Feature Type	# Instances	# Features
Real	150	4

Dataset Information

What do the instances in this dataset represent?
Each instance is a plant

Additional Information
This is one of the earliest datasets used in the literature on classification methods and widely used in statistics and machine learning. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are not linearly separable from each other....

SHOW MORE ▾

Has Missing Values?
No

Por su parte, las siguientes imágenes muestran varios ejemplos de problemas que no serían válidos para ser usados en esta práctica, por distintos motivos. En el primer caso, porque las entradas son imágenes y no atributos (*Data Set Characteristics*), en el segundo caso por tener un número de patrones (# *Instances*) muy bajo (16), en el tercer caso porque en la base de datos faltan valores (*Missing Values*), y en el último caso porque el tipo de problema (*Associated Tasks*) no es de

clasificación o regresión.



Volcanoes on Venus - JARtool experiment

The JARtool project was a pioneering effort to develop an automatic system for cataloging small volcanoes in the large set of Venus images returned by the Magellan spacecraft.

Dataset Characteristics	Subject Area	Associated Tasks
Image	Climate and Environment	Classification
Feature Type	# Instances	# Features
-	1	-

Dataset Information ^

Additional Information

The data was collected by the Magellan spacecraft over an approximately four year period from 1990--1994. The objective of the mission was to obtain global mapping of the surface of Venus using synthetic aperture radar (SAR). A more detailed discussion of the mission and objectives is available at JPL's Magellan webpage....

SHOW MORE ▾

Has Missing Values?

Yes



Balloons

Data previously used in cognitive psychology experiment; 4 data sets represent different conditions of an experiment

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Social Science	Classification
Feature Type	# Instances	# Features
Categorical	16	-

Dataset Information ^

Additional Information

There are four data sets representing different conditions of an experiment. All have the same attributes.

a. adult-stretch.data Inflated is true if age=adult or act=stretch...

SHOW MORE ▾

Has Missing Values?

No



Breast Cancer

Donated on 7/10/1988

Breast Cancer Data (Restricted Access)

Dataset Characteristics

Multivariate

Subject Area

Health and Medicine

Associated Tasks

Classification

Feature Type

Categorical

Instances

286

Features

9

Dataset Information



Additional Information

This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary-tumor.)

...

SHOW MORE ▾

Has Missing Values?

Yes



Sales_Transactions_Dataset_Weekly

Donated on 7/15/2017

Contains weekly purchased quantities of 800 over products over 52 weeks. Normalised values are provided too.

Dataset Characteristics

Multivariate, Time-Series

Subject Area

Business

Associated Tasks

Clustering

Feature Type

Integer, Real

Instances

811

Features

-

Dataset Information



Additional Information

52 columns for 52 weeks; normalised values of provided too.

Has Missing Values?

No

A pesar de que esta página de Universidad de California es una de las más conocidas, existen muchas otras bases de datos y repositorios disponibles en internet que podrían ser válidas para realizar esta práctica. Alguno de los posibles recursos sería:

<https://www.kaggle.com/datasets>

<https://datasetsearch.research.google.com>

<https://huggingface.co/datasets>

Nuevamente, es necesario consultar con el profesor el problema seleccionado para que valide si es adecuado para hacer la práctica o no.

Para la resolución de este problema se ejecutarán las funciones de la práctica 1 necesarias para experimentar con el *dataset* seleccionado, como *one-hot-encoding*, normalización, o validación cruzada. El objetivo es realizar un estudio comparativo de las técnicas estudiadas en prácticas. Como resultado, se tendrán los valores de la(s) métrica(s) seleccionadas para cada configuración de cada algoritmo, lo que permitirá elaborar tablas comparativas de resultados. Es importante establecer la métrica que se utilizará para evaluar los sistemas resultantes, que es dependiente del problema escogido. Por lo tanto, los valores para realizar la comparación serán los resultados de una validación cruzada.

En algunas ocasiones, los datos para descargar vienen separados, incluyendo un conjunto de denominado *test* que sirve para probar los modelos generados, con independencia de los datos usados en la creación del modelo. En estos casos, para hacer que el trabajo de todos los equipos sea similar, lo mejor es ignorar el conjunto de *test* disponible, y realizar validación cruzada con los datos de entrenamiento.

A finales de cuatrimestre se solicitará una entrega final con todo el material desarrollado, en el cual tendrá que figurar estos 5 archivos como mínimo:

- Archivo de funciones del sistema de AA, con las funciones desarrolladas en la práctica 1. Aunque ya se entregó en la práctica anterior, se permite entregar nuevamente (sin modificar la nota de la práctica 1) por si se ha necesario modificar alguna función por fallo que pudiera tener.
- Base de datos usada.
- Archivo de índices con los índices utilizados en la validación cruzada, para repetir los experimentos.
- Archivo ejecutable. La idea es que el profesor de prácticas pueda ejecutar este archivo y que se muestren las tablas de resultados y las gráficas que pudiera haber en la memoria. Para ello, se debería fijar la semilla aleatoria y utilizar el archivo de índices especificado en el punto anterior para hacer validación cruzada.

- Memoria en formato pdf.

Tras comprobar que el código proporcionado puede repetir los experimentos, la evaluación de esta práctica se realizará mediante la memoria entregada. Con respecto a esta, en otro documento se detalla cómo debería ser su redacción. La memoria constituirá el centro de la evaluación del trabajo, puesto que cada parte se debería ver plasmada en ella. A continuación, se especifican los criterios que se utilizarán para corregir las memorias. La puntuación de cada parte de la memoria será la siguiente:

- Introducción: 0.2 puntos. Se valorará la claridad de la explicación.
- Descripción del problema: 0.2 puntos. Se valorará la claridad de la explicación y de los detalles aportados sobre los datos con los que se trabajará. Una explicación que es necesaria en esta parte es la justificación de qué métrica o métricas se utilizarán para valorar y comparar los clasificadores que se generen, y por qué son más adecuadas que el resto de métricas.
- Análisis bibliográfico: 0.2 puntos. Se valorará el número de trabajos descritos, además de las propias descripciones. En este aspecto, para obtener la nota máxima será necesario incluir al menos 8 trabajos y describirlos brevemente. En ocasiones no será posible encontrar referencias directamente relacionadas con la temática a desarrollar, en estos casos se podrán analizar trabajos con temáticas similares. Estos trabajos deberán estar correctamente referenciados siguiendo un estilo concreto, siendo estas referencias de alguna publicación en libros, revistas, actas de congreso o publicaciones similares, pero no de páginas web.
- Desarrollo: 1.5 puntos, que se distribuirán de la siguiente manera:
 - Descripción: 0.45 puntos. Se valorará la claridad de la descripción, de esta parte, incluyendo conceptos como:
 - Descripción de la base de datos utilizada. A pesar de que en la sección “Descripción del problema” ya se hayan descrito los datos, es posible que para su aplicación estos varíen ligeramente, por ejemplo, eliminando atributos irrelevantes, por lo que es conveniente tener una descripción incluyendo cuántos patrones se han usado, cuántas entradas, clases, etc.
 - Preprocesado de los datos, que suele ser relativo a la normalización de estos. Es necesario justificar el porqué del tipo de normalización utilizado,

así como los parámetros de normalización (mínimo, máximo, media, etc.), o por qué no se realiza normalización.

- Otros datos relativos a los experimentos que se van a llevar a cabo, como metodología, número de *folds*, etc.
 - Cualquier otro material como gráficas en las que se muestren los datos puede ser de interés para esta parte.
- Resultados: 0.6 puntos. Esta sección contiene la parte experimental. En ella, se valora la claridad de las explicaciones, así como el número de experimentos. Es necesario realizar experimentación con 5 técnicas vistas en clase (Redes de Neuronas, SVM, Árboles de Decisión, kNN y DoME), y, para cada una, probar con distintos hiperparámetros.
- Para el caso de RR.NN.AA., probar al menos 8 arquitecturas distintas, entre una y 2 capas ocultas.
 - Para SVM, probar con distintos *kernels* y valores de C. Como mínimo, 8 configuraciones de hiperparámetros de SVM.
 - Para DoME, probar al menos 8 valores distintos de número de nodos, en el intervalo donde se hayan obtenido los mejores resultados.
 - Para Árboles de Decisión, probar al menos 6 valores de profundidad distintos.
 - Para kNN, probar al menos 6 valores de k distintos.

En todos los experimentos realizados, es necesario usar las métricas descritas en la sección 2 de la memoria para evaluar y comparar los modelos obtenidos.

- Discusión: 0.45 puntos. En esta sección se valorará la claridad de la explicación y los razonamientos empleados, indicando el impacto que tienen en el sistema desarrollado, que deberán apoyarse en los resultados obtenidos, así como en otros que se quieran mostrar aquí, como pueden ser matrices de confusión o distintas gráficas explicativas. Un punto a tener en cuenta es que se pueden integrar test estadísticos para reforzar los argumentos.
- Conclusiones: 0.2 puntos. Se valorará la claridad de la explicación, así como que las

conclusiones que se extraigan estén apoyadas por los resultados obtenidos.

- Trabajo futuro: 0.2 puntos. Se valorará la claridad de la explicación.