Tarea 2
INSTRUCCIONES

Estas pruebas están basadas en la filosofía de aprender haciendo y la intención es que tratando de realizar los problemas, ejercicios o análisis, los estudiantes ejerciten e interioricen los conocimientos teóricos que han adquirido mediante el estudio de los textos, los tutoriales y los materiales complementarios. Al tratarse de una prueba evaluable se trata de un trabajo individual, es de este modo, pensando los problemas por su cuenta e investigando, como lograrán una buena comprensión de la materia.

Entrega: Las soluciones a los problemas han de ser redactados en un documento realizado mediante una aplicación informática que permita incluir las fórmulas y los desarrollos de los problemas. Existen múltiples opciones como el editor de ecuaciones de Microsoft Word Office, o software libre como LibreOffice. También es posible redactarlas empleando LaTeX. Es recomendable este modo ya que ofrece la mejor presentación para textos científicos y su aprendizaje será de enorme utilidad para aquellos que desarrollan una carrera en Ciencias e Ingeniería. Si deciden tomar esta opción en el curso virtual hay un enlace con un excelente tutorial de LaTeX y una plantilla que pueden usar si lo desean.

La empresa Makichan ha puesto recientemente la posibilidad de acceder de forma gratuita al software Scientific Word. Se trata de un editor de LaTeX que facilita enormemente el manejo de ecuaciones:

https://www.mackichan.com/techtalk/v60/FreeSW.htm

En ningún caso se permitirá la entrega de documentos redactados a mano y escaneados.

A la plataforma debe subirse un archivo .zip .rar .tar o equivalente en el que estén contenidos la memoria en pdf con el formato adecuado y también la fuente: el archivo .doc o .tex o equivalente del que se ha obtenido el pdf. También debe subir el archivo de LiveScript de Matlab.mlx con el que ha realizado los cálculos y exportado a pdf.

Trabajo individual: Se trata de una prueba calificable y por tanto individual. No se permité el intercambio de información entre los estudiantes, que además iría en detrimento de su aprendizaje. El equipo docente, como es su obligación, velará porque esto sea así en la corrección y utilizando las herramientas antiplagio de las que dispone la institución. En caso de ser detectado plagio se verá obligado a actuar al respecto tomando las medidas pertinentes sobre todos los agentes implicados.

1 ESTUDIO DE CONJUNTOS DE DATOS

Esta segunda tarea está enfocada en el tratamiento en detalle de conjuntos de datos. Se trabajará con dos conjuntos de datos diferentes que deben ser analizados por separado. De ellos debe obtener la información que considere pertinente o interesante implementando las diferentes técnicas que se han estudiado hasta el momento: métodos estadísticos, regresión lineal y logística, clusterización, reglas de asociación, etc... No todas las técnicas son de utilidad en todos los problemas y conjuntos de datos e incluso puede carecer de sentido aplicarlas.

Los dos conjuntos de datos que se van a estudiar son clásicos como ejercicios y ejemplos en Ciencia de Datos y Aprendizaje Automático, es posible que encuentre en la red mucha información al respecto y es libre de estudiarla y aprender de ella, pero no es válido reproducirla literalmente. Debe aportar sus propias soluciones y deben justificar las técnicas que usen en función de los resultados que quieran obtener o la información que crean que resulta interesante. Aunque hay ciertos conjuntos de datos -como los que nos ocupan en este ejercicio- que llevan mucho tiempo en circulación y han sido muy estudiados es habitual que de ellos se extraiga nueva información, muchas veces aplicando nuevas técnicas o realizando de forma inteligente transformaciones de los datos, seleccionando subconjuntos interesantes, etc. La información que se puede obtener de un conjunto de datos no está directamente relacionada con el tamaño del mismo. En esta práctica veremos un conjunto de datos pequeño del que se ha extraído mucha información y otro conjunto órdenes de magnitud más grande. La información obtenible depende de la calidad de los datos, lo representativos que sean del problema a tratar y la habilidad del científico de datos para aplicar las técnicas adecuadas y llegar a conclusiones con los resultados obtenidos.

Tenga en cuenta que puede resultar interesante hacer transformaciones de los datos, por ejemplo, obtener una nueva columna que sea el ratio entre dos cantidades, su producto o su suma.

En las partes en las que realice ejercicios de clasificación debe tener en cuenta que en la tarea anterior programó medidas de aciertos y fallos en la regresión logística, como se indicó, son de suma importancia ya que su uso se extiende a todo problema de clasificación. En estos ejercicios debe implementar, donde sea necesario, medidas equivalentes para medir la eficiencia de los modelos que ha propuesto e implementado.

Todos los cálculos y transformaciones deben ser realizados mediante el software Matlab.

Finalmente debe redactar una memoria en la que debe seguir el esquema de presentar una introduc-

ción, luego un desarrollo y finalmente unas conclusiones.

Un esquema posible para la memoria puede ser:

Introducción: Donde se presenta el problema, los datos disponibles y lo que se pretende averiguar.

Tratamiento de datos/Resultados: Donde se explican las técnicas que se usan, y los resultados que se obtienen. De forma crítica por que se escogen unas y no otras, como los resultados de alguna de ellas motivan la elección de otras, etc.

Conclusiones: Un resumen de las conclusiones más importantes que se han obtenido mediante el tratamiento de los datos. Las conclusiones no deben ser muchas, no más de cuatro o cinco y cada una de ellas debe, en unas pocas líneas, explicar un resultado importante.

En muchos trabajos académicos es bastante habitual que lo último que se escriba sea la introducción: durante la redacción del tratamiento de datos y resultados se va perfilando que es lo más importante del trabajo realizado y que conclusiones se pueden sacar, posteriormente se pueden escribir estas y finalmente la introducción que de una idea del trabajo con el que se va a encontrar el lector.

La memoria de cada uno de los problemas a tratar **no debe ocupar más de 10 páginas en total**, incluyendo las figuras: pueden ser incluso necesarias menos. No se trata de hacer un trabajo largo, sino de aprender a seleccionar que técnicas usar y que resultados es importante poner en relevancia en el trabajo. Seguramente haya muchos cálculos de prueba que no merezca la pena incluir en la memoria o resultados menores que sólo merezcan ser comentados en unas líneas, pero que no requieran la atención de aparecer en figuras, tablas, etc. No se trata de hacer el mayor número posible de cálculos y aplicar una gran cantidad de técnicas sino de ver cuáles son interesantes y se valorará la capacidad de síntesis al exponer los resultados y la justificación de utilizar un método u otro. Se debe tener una actitud crítica.

Consulte el documento de orientaciones para la redacción de las tareas donde encontrará indicaciones sobre diversos detalles, formato, etc.

Le resultará también útil la Guía para la Redacción de Trabajos Académicos que la Biblioteca de la UNED ha confeccionado:

https://uned.libguides.com/trabajos_academicos

Los conjuntos de datos están disponibles en el curso virtual ya que se han incluido en algunos ejemplos de los tutoriales de Matlab del material complementario y el conjunto Iris dataset viene incluido en Matlab.

1. 1 Conjunto de datos 1: Iris Dataset.

El conjunto de datos de Iris Dataset es un conjunto de datos relativamente pequeño con apenas 150 entradas. Contiene información de características físicas de varias de flores Iris que se clasifican en tres variedades diferentes:

Versicolor

Setosa

Virginica

Contiene 50 entradas de cada tipo de flor.

Aparte de la variedad a la que pertenecen están registradas cuatro características físicas como son el largo y el ancho tanto de los pétalos como de los sépalos.

El conjunto de datos fue recopilado por el botánico Edgar Anderson:

Edgar Anderson (1936). "The species problem in Iris". Annals of the Missouri Botanical Garden. 23 (3): 457–509. doi:10.2307/2394164. JSTOR 2394164.

Edgar Anderson (1935). "The irises of the Gaspé Peninsula". Bulletin of the American Iris Society. 59: 2–5.

Posteriormente el botánico y estadístico R. A. Fisher realizó un estudio pormenorizado utilizando las técnicas estadísticas disponibles en la época:

R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics. 7 (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x. hdl:2440/15227.

Es un conjunto de datos que se ha usado como modelo en problemas de separación de las variedades en función de las características físicas pero que también puede tener interés desde el punto de estudiar si tales características tienen algún tipo de relación dependiendo de la especie, etc.

Hay total libertad para experimentar con los datos y obtener información que considere relevante del conjunto o determinar si unas técnicas funcionan mejor que otras para un determinado propósito, etc.

1. 2 Conjunto de datos 2: Diamonds Dataset.

El conjunto de datos de Diamonds Dataset es un conjunto de datos de tamaño mediano que contiene información de 50000 diamantes. Cada una de las entradas contiene las siguientes características:

carat: peso del diamante en kilates

cut: calidad del corte del diamante

color: calidad del color del diamante desde D que es el mejor a J que es el peor.

clarity: medida de la claridad del diamante de mejor a peor: I1 , SI2, SI1, VS2, VS1, VVS2, VVS1, IF.

table: anchura en la parte alta del diamante

x: longitud -eje x- en milímetros.

y: anchura -eje y- en milímetros

z: profundidad -eje z- en milímetros

depth: profundidad total expresada en porcentaje, obtenida mediante la fórmula: depth=2*z/(x+y)

price: precio en dólares estadounidenses

Este conjunto de datos también ha sido muy estudiado y por la gran cantidad de características es muy adecuado para encontrar correlaciones o dependencias funcionales, etc.

Hay total libertad para experimentar con los datos y obtener información que considere relevante del conjunto o determinar si unas técnicas funcionan mejor que otras para un determinado propósito, etc.