PEC1

Presentación

Primera actividad de evaluación continúa del curso. En esta PEC se practicaran los algoritmos básicos de categorización.

Competencias

Competencias de grado

- Capacidad de utilizar los fundamentos matemáticos, estadísticos y físicos y comprender los sistemas TIC.

- Capacidad para analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.

- Capacidad para conocer las tecnologías de comunicaciones actuales y emergentes y saberlas aplicar convenientemente, para diseñar y desarrollar soluciones basadas en sistemas y tecnologías de la información.

- Capacidad para proponer y evaluar diferentes alternativas tecnológicas y resolver un problema concreto.

Competencias específicas

- Capacidad para utilizar la tecnología de aprendizaje automático más adecuada para resolver un determinado problema.

- Capacidad para evaluar el rendimiento de los diferentes algoritmos de resolución de problemas mediante técnicas de validación cruzada.

Objetivos

El objetivo de esta prueba de evaluación es categorizar los datos de los archivos adjuntos relacionados con billetes falsificados. Queremos agrupar los datos de distintos billetes en función de si son auténticos o bien son falsificaciones.

Descripción de los datos

Los archivos proporcionados (LARGE.CSV y SMALL.CSV) tienen un formato tipo tabla en el que cada fila es un ejemplo. Las 4 primeras columnas son atributos y

la quinta columna representa la clase. El separador de columnas utilizado es la coma (,) y el separador decimal de números es el punto (.). La primera fila no contiene datos puesto que es la cabecera.

Los datos que contienen los archivos provienen del *Banknote Authentication dataset* disponible en:

https://archive.ics.uci.edu/dataset/267/banknote+authentication

Los atributos de los archivos proporcionados (4 primeras columnas) se corresponden con los atributos del mencionado *Banknote Authentication dataset*. Dichos atributos provienen de un procesado de imagen de los billetes, y podéis consultar su significado en el enlace anterior. La quinta columna (clase) de los archivos proporcionados puede valer 0 si el billete es una falsificación o 1 si el billete es auténtico.

Nótese que, aunque se puede (y se recomienda) consultar el significado de los atributos en el enlace anterior, no se debe trabajar con los archivos originales del dataset. Debéis trabajar con los que os proporcionamos: LARGE.CSV y SMALL.CSV. En cada ejercicio se os indicará con cuál de los dos archivos (LARGE.CSV o SMALL.CSV) debéis trabajar.

Descripción de la PEC

Conjuntamente con esta PEC se os proporcionan dos programas en Python (kmeans_sklearn.py y pca_basic.py) por si os pueden ser de utilidad. Estos programas no necesariamente funcionarán directamente con los datos suministrados y el tratamiento previo de los datos que se llevan a cabo a modo de ejemplo en ellos no necesariamente es el adecuado para esta PEC. Los programas sólo se proporcionan como referencia. Se recomienda leer el apartado "Criterios de Valoración" para saber como se evaluará la prueba.

Ejercicio 1

Para realizar este ejercicio, no se debe utilizar un programa que lo resuelva; se espera que mostréis y justifiquéis cada uno de los pasos y cálculos realizados.

En este ejercicio, debéis comprobar si es posible categorizar el archivo de datos pequeño (SMALL.CSV). En particular, se solicita:

a) Efectuad, si es necesario, el tratamiento previo de los datos. Justificad todas las decisiones que toméis. Mostrad el conjunto de datos tratado. En los apartados posteriores, cuando se hable de "SMALL.CSV", se entenderá que se refiere al resultado del tratamiento de datos resultante de este apartado.

b) Utilizad K-Means nítido para categorizar los datos de dicho archivo en dos categorías, ignorando las columnas no pertinentes. ¿Cuál es el nivel de exactitud del resultado? Nótese que el término "exactitud" se refiere a lo que, en inglés, se denomina “accuracy”, no a “precision”.

c) Aplicad el algoritmo PCA para reducir la dimensionalidad del conjunto anterior conservando el 95% de la varianza. Utilizad K-Means nítido sobre el conjunto reducido de la misma forma que en el apartado anterior. Comparad los resultados.

Ejercicio 2

En este ejercicio construiréis dendrogramas para categorizar los datos de SMALL.csv. Nótese que se debe trabajar con los datos tratados en el ejercicio anterior, no con los datos originales ni tampoco con los resultados de aplicar PCA. Se pide:

a) Construid el dendrograma utilizando el enlace simple. Mostrad todos los pasos implicados, incluyendo el dendrograma resultante y explicad las decisiones tomadas. Indicad también muy claramente cuál sería la categorización propuesta por el dendrograma y calculad la exactitud.

b) Construid el dendrograma utilizando el vínculo completo. Mostrad todos los pasos implicados, incluyendo el dendrograma resultante y explicad las decisiones tomadas. Indicad también muy claramente cuál sería la categorización propuesta por el dendrograma y calculad la exactitud.

c) Construid el dendrograma utilizando el vínculo medio (average). Mostrad todos los pasos implicados, incluyendo el dendrograma resultante y explicad las

decisiones tomadas. Indicad también muy claramente cuál sería la categorización propuesta por el dendrograma y calculad la exactitud.

Ejercicio 3

En este ejercicio se trabajará con el archivo de datos grande (LARGE.csv) y se utilizará la biblioteca sklearn además de, probablemente, NumPy. Podéis encontrar información en http://scikit-learn.org . Quizá debáis instalarla en vuestro ordenador aunque es probable que ya lo esté (muchas distribuciones de Python ya la incluyen y, por ejemplo, Google Colab ya dispone de ella).

a) Cargad los datos de LARGE.CSV en un array NumPy y responded a las siguientes preguntas: ¿cuántos ejemplos hay? ¿de cuántos atributos consta cada ejemplo? ¿cuántas clases distintas existen? ¿cuántos ejemplos de cada clase hay? Mostrad el código que proporciona los valores solicitados. Notad que de algunas preguntas (número de clases o número de atributos) ya conocéis la respuesta. Lo importante en este apartado no es tanto la respuesta (que debe ser correcta) como que mostréis el código que la proporciona.

b) Aplicad a los datos de LARGE.CSV el mismo tratamiento que se aplicó en el ejercicio 1. Mostrad el código que efectúa dicho tratamiento así como los 10 primeros ejemplos tratados (mostrad tanto los atributos tratados como su clase).

c) Aplicad sobre los datos tratados (todos ellos, no sólo los 10 ejemplos que hayáis mostrado) y utilizando sklearn el algoritmo K-Means con dos ejemplos escogidos manualmente como centroides iniciales. Indicad cuál es la exactitud (accuracy), la precisión (precision), el recall, el fall-out y la matriz de confusión. Mostrad el código que lleva a cabo la tarea solicitada.

d) Repetid el apartado anterior ahora utilizando el algoritmo K-Means++ para la selección de los centroides iniciales.

Recursos

Básicos

Para realizar esta PEC disponéis de los archivos adjuntos ("SMALL.CSV" y "LARGE.CSV") donde encontraréis los datos con los que deberéis trabajar. También os adjuntamos dos ejemplos en Python ('kmeans_sklearn.py' y

'pca_basic.py'). Nótese que estos ejemplos pueden no funcionar directamente con los datos proporcionados. Si queréis utilizarlos como soporte para resolver la PEC deberéis revisarlos cuidadosamente y parametrizarlos de forma adecuada. En particular, el tratamiento de datos que se lleva a cabo en estos archivos se proporciona únicamente a modo de ejemplo, por lo que no necesariamente se corresponde con el tratamiento que deberéis hacer vosotros.

Criterios de valoración

Los tres ejercicios de esta PEC se valoraran con 10/3 puntos. Cada ejercicio se valorará globalmente, de forma que no existe una puntuación específica para cada apartado.

En todos los casos, además de los resultados obtenidos, se valorará la explicación del trabajo realizado, las valoraciones aportadas y las justificaciones proporcionadas. Un resultado sin justificar se valorará con cero puntos.

Formato y fecha de entrega

Deberéis entregar la PEC en formato PDF mediante el registro de actividades de evaluación contínua dentro del plazo establecido en el cronograma del aula de la asignatura.

Para dudas o aclaraciones sobre el enunciado, dirigíos al consultor responsable de vuestra aula.

Nota: **Propiedad intelectual**

A menudo es inevitable, al producir una obra multimedia, hacer uso de recursos creados por terceras personas. Es por tanto comprensible hacerlo en el marco de una práctica de los estudios del Grado de Informática, siempre que esto se documente claramente y no suponga plagio en la práctica.

Por lo tanto, al presentar una práctica que haga uso de recursos ajenos, se presentará junto con ella un documento en el que se detallen todos ellos, especificando el nombre de cada recurso, su autor, el lugar donde se obtuvo y el su estatus legal: si la obra está protegida por copyright o se acoge a alguna otra licencia de uso (Creative Commons, licencia GNU, GPL ...). El estudiante deberá asegurarse de que la licencia que sea no impide específicamente su uso en el marco de la práctica. En caso de no encontrar la información correspondiente deberá asumir que la obra está protegida por copyright.

Deberán, además, adjuntar los archivos originales cuando las obras utilizadas sean digitales, y su código fuente si corresponde.