PAC 2: Extracció y selección de atributos`

Presentación

En esta prueba se aplicarán diferentes técnicas de selección de atributos a un problema real: un conjunto de dígitos escaneados escritos a mano por diferentes personas. A estas imágenes se les ha hecho un preprocesado previo para que el estudiante pueda centrarse en las técnicas de reducción de la dimensionalidad.

Competencias

En este enunciado se trabajan en un determinado grado las siguientes competencias general de máster:

- Capacidad para proyectar, calcular y diseñar productos, procesos e instalaciones en todos los ámbitos de la ingeniería en informática.
- Capacidad para el modelado matemático, cálculo y simulación en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación, desarrollo e innovación en todos los ámbitos relacionados con la ingeniería en informática.
- Capacidad para la aplicación de los conocimientos adquiridos y para solucionar problemas en entornos nuevos o poco conocidos dentro de contextos más amplios y multidisciplinares, siendo capaces de integrar estos conocimientos.
- Poseer habilidades para el aprendizaje continuado, autodirigido y autónomo.
- Capacidad para modelar, diseñar, definir la arquitectura, implantar, gestionar, operar, administrar y mantener aplicaciones, redes, sistemas, servicios y contenidos informáticos.
- Capacidad para asegurar, gestionar, auditar y certificar la calidad de los desarrollos, procesos, sistemas, servicios, aplicaciones y productos informáticos.

Las competencias específicas de esta asignatura que es trabajan son:

- Entender que es el aprendizaje automático en el contexto de la Inteligencia Artificial.
- Distinguir entre los diferentes tipos y métodos de aprendizaje.
- Aplicar las técnicas estudiadas en un caso concreto.

Objetivos

En esta PAC s'aplicaran a un cas concret els conceptes del temari sobre extracció i selecció d'atributs (tema 3).

Descripción de la PEC a realizar

Datos

El conjunto de datos consta de 2000 dígitos del 0 al 9 escritos a mano por diferentes personas. Las imágenes obtenidas han sido previamente limpiadas de ruido, centradas y escaladas (15x16 píxeles), de manera que todas tienen el mismo tamaño.

El número de dígitos de cada clase 0..9 es el mismo (200), y todos los archivos de datos están ordenadas (los primeros 200 son '0', los siguientes 200 son '1' , etc.).

En este conjunto de datos se proporcionan diferentes medidas útiles para el procesado de imágenes, de las que en este caso nos interresan las siguientes:

Los píxeles promediados en ventanas de 2x3 (240 valores, archivo mfeat-pix.txt) .

Las correlaciones de los perfiles (216 valores, archivo mfeat-fac.txt) .

Todos los apartados de esta práctica se aplicarán separadamente a cada uno de los ficheros de datos anteriores, y se compararán los resultados obtenidos en ambos conjuntos de características .

Este conjunto de datos se llama "Multiple Features" y ha obtenido del Machine Learning Repository de la Universidad de California Irvine :

https://archive.ics.uci.edu/ml/datasets/Multiple+Features

Ejercicio 1

Efectuar, si es necesario, el tratamiento previo de los datos (tratamiento de valores ausentes o erróneos, normalización ,centrado , etc .). Justificación de la necesidad de cada operación según las características del problema.

Para poder aplicar los métodos requeridos en esta PEC, hay que normalizar los datos, restar la media de cada variable (cada columna) i dividir por la desviación estándar. También hay que comprobar si hay valores ausentes o erróneos, cosa que en este caso no se da.

Todo el código que resuelve esta PEC está en el fichero pac2.py, con cada apartado indicado con comentarios.

Ejercicio 2

Aplique un análisis PCA a los datos resultantes del ejercicio anterior y estudia los vectores propios y las varianzas resultantes. Debe utilizar como modelo el código 3.5 de los materiales. Cuántos componentes son necesarios para obtener un 95% de varianza?.

Los valores propios resultantes de aplicar PCA indican la cantidad de varianza de los datos originales que explica cada componente principal. En este caso se ha utilizado la función PCA de **skearn**, que da como uno de sus resultados el tanto por uno de varianza explicada (es decir, la varianza de cada componente dividida por la suma de todas las varianzas). Así, sumando estos valores se puede ver en qué momento se llega a acumular el 95% de la varianza total de los datos.

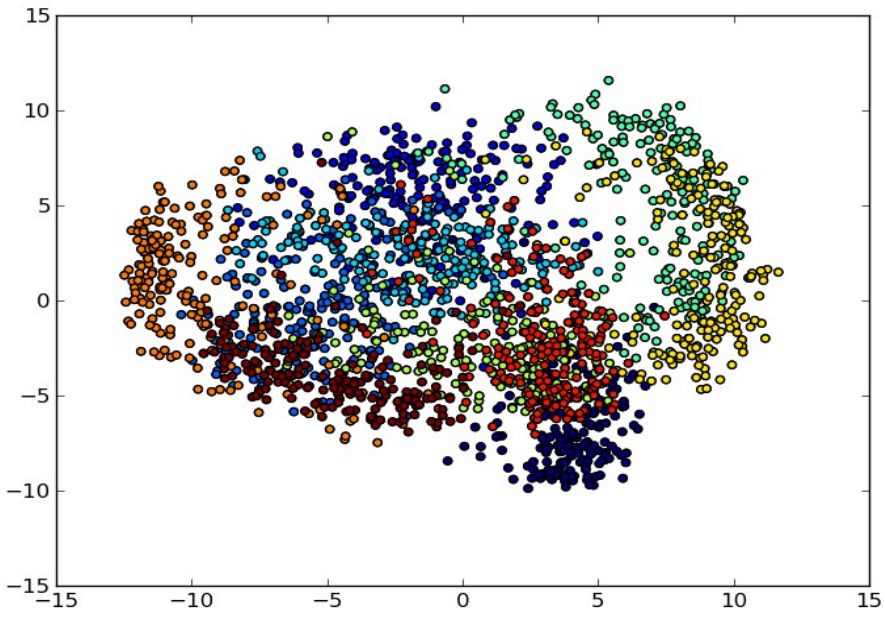
Aplicando este cálculo a los datos de píxels, se obtiene que se necesitan 89 componentes para explicar el 95% de varianza; pero cuando se aplica a los factores de correlación el número de componentes se reduce a 13. Esto se debe a que la información que dan los píxels es muy difusa, no hay píxels (o combinación de ellos) que claramente diferencien las distintas imágenes (independientemente de su clase).

Pero como los factores de correlación son una medida más elaborada que tiene en cuenta más información, con pocos componentes se resumen bastante bien las diferencias entre muestras, porque las variables de entrada son más ricas por lo que respecta a su carga informativa.

Ejercicio 3

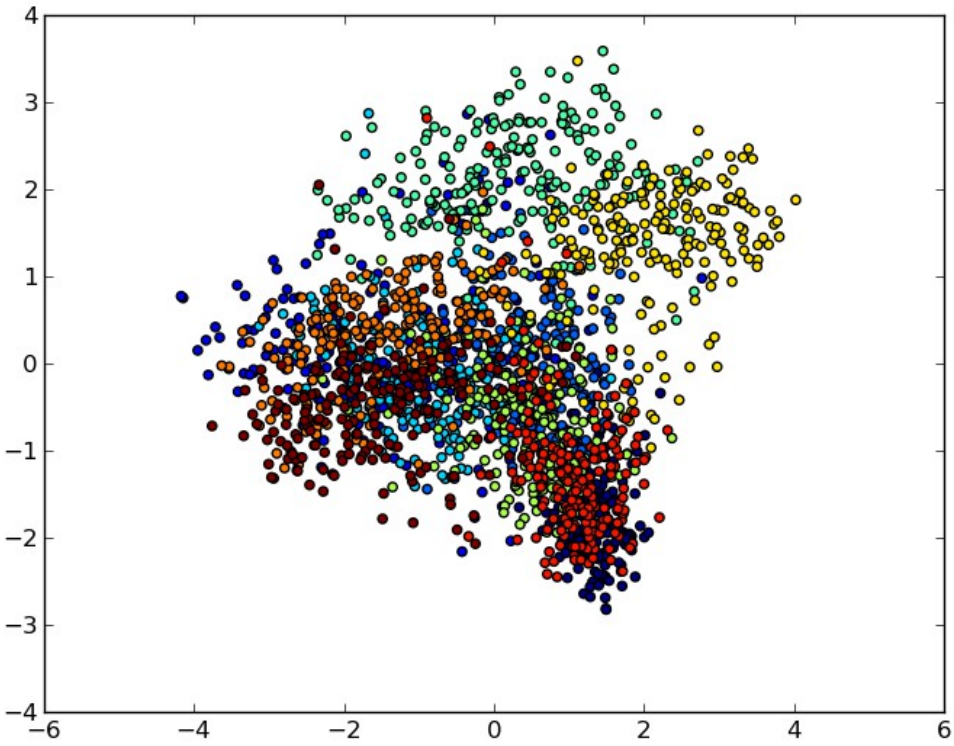
Con los resultados del ejercicio anterior, proyectar los datos con dos componentes. Al hacer la gráfica dibuja los puntos de cada una de las clases de un color diferente. ¿Qué conclusiones se pueden sacar?.

En primer lugar podéis ver la proyección de los dígitos según sus dos primeras componentes principales aplicadas a los datos de píxels. Se puede observar que hay una cierta agrupación de muestras por colores (cada colores una cifra diferente del 0 al 9). Parece que será posible clasificar las muestras de acuerdo con esta información, aunque la separación dista de ser perfecta.



Por otra parte, cuando se proyectan los datos del fichero de factores de correlación, se obtiene la imagen siguiente, en la que la separación entre algunas clases es mejor (amarillo, rojo, verde), pero también hay más concentración de muestras en el centro de la gráfica.

Posiblemente cada una de las proyecciones separa mejor unas clases que otras. De hecho eso es exactamente lo que ocurre, pues el objetivo de este conjunto de datos es precisamente utilizar diferentes vistas de los mismos datos para mejorar la clasificación, teniendo en cuenta que cada vista ayuda en determinados casos.



`

Ejercicio 4

Aplique LDA a los datos resultantes del ejercicio 1. Tome como modelo el código 17.3 de los materiales. Determine el porcentaje de acierto al distinguir unos dígitos de otros. Interprete los resultados obtenidos y compárerlos con los resultados obtenidos de los apartados anteriores.

En este apartado se entrena un clasificador LDA sobre los datos del problema para determinar hasta qué punto se podrían clasificar los datos. La función LDA da una medida de la calidad de la clasificación (**score**) como la

proporción de clasificaciones correctas (también se podría aplicar predict y ver cuántas muestras se han clasificado correctamente).

El grado de acierto por píxeles es 0,9775, y por factores de correlación de 0,9923. Son valores muy altos, pero hay que tener en cuenta que estos resultados se han obtenido entrenando y probando sobre los mismos datos. Para hacer una evaluación más estricta de un clasificador, como se verá más adelante, hay que dividir los datos en entrenamiento y prueba.

En todo caso, estos resultados indican que es posible clasificar los datos (unos valores bajos en este punto nos tendrían que hacer plantear si seguir adelante con estas variables de entrada o elegir otras mejores).

Recursos

Este PEC requiere de los siguientes recursos:

Bàsicos: Archivos de datos mfeat-fac.txt y mfeat-pix.txt, que también pueden obtenerse desde:

https://archive.ics.uci.edu/ml/datasets/Multiple+Features

Complementarios: manual de teoria de la asignatura y lcódigo Python del capítulo 3 (en el tablón de la asignatura).

Criterios de valoración

Los ejercicios tendrán la siguiente valoración asociada:

- Ejercicio 1: 1 punto

- Ejercicio 2: 3 puntos. Se valorará con 2 puntos si se calcula correctamente el PCA y los valores y vectores propios. El cálculo de la dimensionalidad que alcanza el 95% de la varianza vale 1 punto.

- Ejercicio 3: 3 puntos. Se valorará con 1 punto la representación gráfica básica, con 1 punto la representación de las clases (con colores), y con 1 punto las conclusiones.

- Ejercicio 4: 3 puntos. Se valorará con 1 punto la aplicación de LDA, con 1 punto el cálculo del porcentaje de acierto, y con 1 punto las conclusiones y comparación con apartados anteriores.

Es necesario razonar las respuestas en todos los ejercicios. Las respuestas sin justificación no recibirán puntuación.

Formato y fecha de entrega

La PEC debe entregarse antes del **próximo 24 de Abril** (antes de las 24h).

La solución a entregar consiste en un informe en formato PDF usando la plantilla colgada en el tablón de la asignatura más los archivos de código (*. Py) que usó para resolver la prueba. Estos archivos deben comprimir en un archivo ZIP.

Adjuntar el fichero a un mensaje en el apartado de **Entrega y Registro de AC (RAC)**. El nombre del archivo debe ser ApellidosNombre_IA_PEC1 con la extensión. zip.

Para dudas y aclaraciones sobre el enunciado, diríjase al consultor responsable de su 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 en Informática, siempre y 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.