PRAC2: LECTURA DE BASES DE DATOS

Asignatura: Clasificación de Patrones
ETSETB
UPC

UPC-TSC-D5

Octubre - 2016

1 Objetivos de la práctica 2

- Utilizar una de las bases de datos reales (no sintéticas) obtenidas a través del siguiente link:

  http://www-stat.stanford.edu/~tibs/ElemStatLearn/data.html

- Reducir la dimensión de las bases de datos atendiendo a criterios PCA.

2 Base de datos: Phoneme

2.1 Características de la base de datos

Esta es la información que los autores de la base de datos facilitan sobre la misma:

```
These data arose from a collaboration between  Andreas Buja, Werner
Stuetzle and Martin Maechler, and we used as an illustration in the
paper on Penalized Discriminant Analysis by Hastie, Buja and
Tibshirani (1995), referenced in the text.

The data were extracted from the TIMIT database (TIMIT
Acoustic-Phonetic Continuous Speech Corpus, NTIS, US Dept of Commerce)
which is a widely used resource for research in speech recognition.  A
dataset was formed by selecting five phonemes for
classification based on digitized speech from this database.  The
phonemes are transcribed as follows: "sh" as in "she", "dcl" as in
"dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and
"ao" as the first vowel in "water".  From continuous speech of 50 male
speakers, 4509 speech frames of 32 msec duration were selected,
approximately 2 examples of each phoneme from each speaker.  Each
speech frame is represented by 512 samples at a 16kHz sampling rate,
and each frame represents one of the above five phonemes.  The
breakdown of the 4509 speech frames into phoneme frequencies is as
follows:

  aa   ao dcl   iy  sh
 695 1022 757 1163 872

From each speech frame, we computed a log-periodogram, which is one of
several widely used methods for casting speech data in a form suitable
for speech recognition.  Thus the data used in what follows consist of
4509 log-periodograms of length 256, with known class (phoneme)
memberships.

The data contain 256 columns labelled "x.1" - "x.256", a response
column labelled "g", and a column labelled "speaker" identifying the
different speakers.
```

Cada vector se ha construido mediante 256 valores (o características), lo que supone una frecuencia máxima de 8 kHz. En la práctica 2 se trabaja con un número máximo de 64 coordenadas o características, que se corresponde con la parte del espectro entre 0 y 2kHz.

El código del fichero de Matlab, *prac2_fonemas.m* facilitado en esta práctica:

- Realiza la lectura de la base de datos de vectores y selecciona inicialmente 64 características.
- Elimina el vector media de todos los vectores en la base de datos.
- Presenta el dibujo de los diferentes espectros, en este caso se dibuja el espectro completo hasta 8 kHz. Para ello debe estar activada la opción mediante la variable inicial i_dib.
- Divide la base de datos en base de datos de Train (70%) y en base de datos de Test (30%). Observe que la división se realiza de forma aleatoria y manteniendo la proporcionalidad Train/Test por clases.
- Invoca a los clasificadores *LC y QC*.
- Calcula probabilidades de error y matrices de confusión.
- En el caso de trabajar únicamente con dos características presenta un scatter de los vectores y las fronteras de decisión obtenidas para la clase 'aa' respecto a cada una de las cuatro clases restantes.

2.2 Clasificación aplicada a la base de datos Phoneme

A realizar en el laboratorio mediante el programa *prac2_fonemas.m*:

Ejecute el fichero tal como lo ha descargado e incluya en el documento entregable con comentarios pertinentes las siguientes respuestas:

- Los diferentes espectros obtenidos para los diferentes fonemas.
- Las probabilidades del error de clasificación de train y de test, obtenidas mediante *LC* y QC usando todas las características.
- Introduzca las matrices de confusión obtenidas en test y comente el resultado de las mismas.

Suponga ahora que solo pudiera trabajar con 2 características o coordenadas de cada vector debido a costes computacionales. Para buscar 2 coordenadas muy discriminativas puede ayudarse de las gráficas de las medias del espectro por clases y seleccionar las 2 coordenadas aparentemente más útiles para la discriminación por clases a través de la variable *v_coor*.

Para este caso se pide incluya en el documento entregable:

- Las probabilidades del error de clasificación de train y de test, obtenidas mediante *LC* y QC. Compare con el caso anterior.
- Incluya el dibujo del scatter con las fronteras de clasificación obtenidas para la clase 'aa' respecto al resto de clases.

2.3 Reducción de dimensión mediante PCA

En el apartado anterior se ha reducido la dimensión, en función de la observación directa y la información relativa a cada coordenada, evaluada de forma subjetiva. En esta parte se propone reducir la dimensión atendiendo a un criterio de componentes principales, para lo cual puede invocar a la subrutina *pca.m*.

Cree un nuevo programa *prac2_fonemas_PCA.m*. Presente para la base de datos completa FONEMAS (d=64), una gráfica del error obtenido en clasificación con *LC/QC* para la base de datos de train/test al reducir la dimensión mediante la técnica *pca* desde 1 hasta 64 (d'=1:64). Programe de tal modo que las cuatro gráficas de error aparezcan en el mismo "plot" , por ejemplo en distinto color. Tenga en cuenta que tanto para diseñar la matriz de proyección como para diseñar el clasificador sólo debería utilizar la base de train, mientras que la base de test se utiliza únicamente para testear.

Puede revisar el *help* de *figure*, *plot, hold*, *legend, title, grid.*

A la vista de los resultados proponga qué dimensión (mediante *pca*) sería la más adecuada para clasificar usando *LC* y cuál sería la más adecuada para clasificar mediante *QC*. Al responder observe que no se debe producir sobre-entrenamiento. Es decir, el error de test no debe ser excesivamente mayor que el error de train.

Incluya en el documento a entregar:

- Código Matlab generado en esta parte.
- Gráfica de los errores de clasificación y comentarios pertinentes según el texto anterior.

NOTA IMPORTANTE

Entregue por Atenea un único documento G***_apellido1_apellido2_prac2.pdf.