PRAC3: SELECCIÓN DE CARACTERÍSTICAS ANÁLISIS DE COMPONENTES PRINICIPALES Y ANÁLISIS POR MÚLTIPLES DISCRIMINANTES

Asignatura: Clasificación de Patrones.
Optativa de Grados
ETSETB
UPC

UPC-TSC-D5
Octubre 2016

Contenido

1. Objetivos de la práctica 3

- Evaluar dos técnicas de selección de características: Análisis de componentes principales (PCA) y análisis por múltiples discriminantes (MDA) (o discriminante de Fisher).
- Calcular umbrales de decisión para tres clases sobre datos unidimensionales (una sola característica).
- Aplicar técnica MDA de reducción de características sobre base de datos reales.

En ambos casos (PCA y MDA) se partirá de la matriz de scatter **S** total sobre el conjunto de datos, y se generará una matriz de transformación **W** única para todas las clases. En el caso PCA, la matriz **W** contiene los autovectores de **S** asociados a los mayores autovalores. En el caso MDA, la matriz **W** contiene los autovectores asociados a los mayores autovalores de $\mathbf{S}_C^{-1}\mathbf{S}_B$. $\mathbf{S}_B$ es la matriz de dispersión entre clases, y $\mathbf{S}_C$ es la suma de las matrices de covarianza de cada clase y mide la dispersión intra-clases.

2. Laboratorio

2.1. Selección de características con bases de datos gaussianas.

Mediante el software suministrado (*prac3_main.m*) se genera una base de datos de entrenamiento y una base de datos de test, con tres clases y vectores de tres coordenadas. Cada una de las clases es Gaussiana, con medias y matrices de covarianza seleccionables en el fichero *prac3_gengauss.m*. Los vectores generados dentro de cada clase son mutuamente independientes. Se calculan los clasificadores lineales y cuadráticos y las probabilidades de error de cada uno de ellos.

A continuación, se seleccionan 2 y 1 características y se clasifican utilizando clasificadores lineales y cuadráticos. Se pretende comparar resultados en función del número de características.

A realizar en el laboratorio:

1. Ejecute *prac3_main.m*. Seleccione PCA, un valor para la semilla aleatoria y SNR=10dB. Valores distintos de la semilla dan lugar a distintos autovectores de las matrices de covarianza de cada clase. En el documento pdf a entregar, incluya una tabla de los errores *LC* y *QC* obtenidos en entreno y en test para cada una de las tres dimensiones.

2. Repita el punto anterior seleccionando el método de MDA para realizar la selección de características y utilizando la misma semilla para generar los mismos datos. Gire la figura “Datos 3D” y compruebe que existen proyecciones 2D en las que los tres clusters están más separados y otras proyecciones en las que las clusters están muy superpuestos.

3. Repita los dos apartados anteriores con SNR=0dB. Compare los scatters en 1D y en 2D obtenidos con respecto a los calculados mediante el discriminante MDA y justifique las diferencias en probabilidad de error al usar MDA o PCA.

A partir de este punto y para comprobar la efectividad de las técnicas anteriores se trabajará con una base de datos gaussiana, tal que los centroides de las 3 clases se hallen alineados y las 3 clases tengan matrices de covarianza idénticas, provocando que la dirección de alineación de las clases corresponda a la de mínima varianza. Para ello cambie la instrucción *prac3_gengauss* por la de *prac3_gengauss_al* en el fichero prac3_main.

4. Dé la expresión de los vectores correspondientes a las 3 medias para una semilla dada e indique el valor de la semilla utilizada ¿Cuál es el rango de la matriz $\mathbf{S}_B$ en este caso? Comente en consecuencia, cuantas características serían estrictamente necesarias usando MDA.
5. Repita de nuevo los puntos iniciales para PCA y MDA con valores de SNR de +5 dB y -5dB. Justifique en que situación MDA resulta claramente ventajoso respecto a PCA.

2.2. MDA en clasificación

En la última parte de la práctica 2 se realizó un análisis de errores de clasificación al aplicar clasificadores lineales y cuadráticos a la base de datos FONEMAS proyectada mediante PCA a una dimensión *d'*=1:256. Ahora se pide que modifique el código de dicha parte para reducir la dimensión de la base de datos mediante MDA. Siga para ello las siguientes etapas:

1.- Ejecute de nuevo el código Matlab de la práctica 2 que analiza los errores de clasificación al proyectar sobre *d'*=1:256 aplicando PCA como método de reducción de características; Incluya en el documento a entregar el código Matlab correspondiente, así como la gráfica de los 4 errores de clasificación obtenidos (train y test mediante LC y mediante QC) en función de *d'*.

2.- Modifique el fichero Matlab aplicando MDA como método de reducción de coordenadas: *d'*=1: d_{MAX} . Para ello justifique en este caso cual será el valor máximo que puede utilizar en la proyección: d_{MAX} . Incluya en el documento a entregar el código Matlab correspondiente, así como la gráfica de los 4 errores de clasificación obtenidos en función de *d'*.

3.- Compare los resultados obtenidos en los 2 apartados anteriores.