PRAC4: K Nearest Neighbors y Ventanas de Parzen

CLP: Clasificación de Patrones.
ETSETB
UPC

Octubre 2016

1 Introducción a la práctica 4

En esta práctica se trabajará con dos bases de datos obtenidas a través del siguiente link:

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

Base ZIP
Cada vector de d=256 muestras representa una imagen de 16×16 pixels de un conjunto de 10 clases diferentes correspondientes a los 10 dígitos (0,…,9) escritos a mano.
Total de vectores de Train: 7291
Total de vectores de Test: 2007
Dimensión: 256
Clases: 10

El programa facilitado internamente combina estas bases de datos (train y test) y genera una nueva base de train con el 50 % de los elementos y una nueva base de test con el 50 % de los elementos. La distribución de elementos entre train y test se realiza de forma aleatoria.

Base MicroArray
Cada vector de d=6830 muestra representa una medida de los genes contenidos en células correspondientes a 14 tipos diferentes de órganos infectados por cáncer.
Total de vectores: 64
Dimensión: 6830
Clases: 14

Se aplicará el método de clasificación KNN, es decir, para clasificar cada nuevo vector se compara con todos los elementos de la base de entreno y a partir de los K vectores de la base más próximos, se decide la clase como la más mayoritaria. Cuando los vectores son de gran dimensión en comparación con el número de vectores de entrenamiento, como nos ocurre con la base de datos *Microarray*, es un método adecuado.

También se probará el método de clasificación basado en ventanas de Parzen y en discriminadores lineales (LC).

2 Base de Datos ZipCode

El ZIP code (Zone Improvement Plan) es el nombre dado en los EEUU al código postal. Esta base de datos puede usarse para OCR (Optical Character Reconigtion).

2.1 Características de la base de datos

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

> Normalized handwritten digits, automatically scanned from envelopes by the U.S. Postal Service. The original scanned digits are binary and of different sizes and orientations; the images here have been deslanted and size normalized, resulting in 16 x 16 gray scale images (Le Cun et al., 1990).

The data are in two zipped files, and each line consists of the digit id (0-9) followed by the 256 grayscale values.

There are 7291 training observations and 2007 test observations, distributed as follows:

	0	1	2	3	4	5	6	7	8	9	Total
Train	1194	1005	731	658	652	556	664	645	542	644	7291
Test	359	264	198	166	200	160	170	147	166	177	2007

or as proportions:

	0	1	2	3	4	5	6	7	8	9
Train	0.16	0.14	0.1	0.09	0.09	0.08	0.09	0.09	0.07	0.09
Test	0.18	0.13	0.1	0.08	0.10	0.08	0.08	0.07	0.08	0.09

Alternatively, the training data are available as separate files per digit (and hence without the digit identifier in each row)

The test set is notoriously "difficult", and a 2.5% error rate is excellent. These data were kindly made available by the neural network group at AT&T research labs (thanks to Yann Le Cunn).

2.2 Compresión de Imágenes

Con el objeto de reducir la dimensión de la base, se aplicarán en esta práctica dos técnicas diferentes de transformación de imágenes: Transformada coseno (DCT) y Transformada de Hadamard (HT).

Una vez en el dominio transformado se reducirá la dimensionalidad de los vectores, pasando de dimensión 256 a 64, a 16 y a 4 y se evaluará el error de clasificación obtenido mediante: LC, KNN y Parzen.

2.3 Laboratorio

2.3.1 Método KNN

1. Ejecute el fichero prac4_zip.m (opción de no transformar imágenes) y observe la imagen representada por los números, tanto para la base de train como para la base de test. Si no se ven bien las imágenes, maximice las figuras.

2. Incluya los errores obtenidos (train, test para LC y KNN) en una única tabla y comente los resultados. Anote el valor del parámetro k que utiliza en KNN. A la vista de las matrices de confusión comente entre que dos clases se cometen más errores.

3. Ejecute de nuevo el fichero prac4_zip.m con transformada DCT y N_dim=16 (en este caso, al no reducir la dimensión es equivalente a lo obtenido en el apartado anterior) y a partir de las representaciones de las transformadas DCT, comente cómo se distribuye la potencia en el dominio transformado y proponga una posible aplicación relacionada con la compresión de imágenes.

4. Para la transformada DCT y para la transformada HADAMARD observe las imágenes resultantes y evalúe de nuevo los errores en las siguientes situaciones.
   N_dim=8 (Matrices de 8×8)
   N_dim=4 (Matrices de 4×4)
   N_dim=2 (Matrices de 2×2)
   Incluya en el documento a entregar una única tabla de errores para los métodos de clasificación KNN y LC, train y test, incluyendo los dos tipos de transformación y las 3 dimensiones propuestas y comente los resultados obtenidos. ¿Qué combinación de las 6 posibles: 2 (DCT, HT) x 3 (N_dim =8, 4, 2) con el método KNN, sería el más apropiado en cuanto a eficiencia computacional, manteniendo el error de test por debajo del 10%?

5. Para la combinación seleccionada en el apartado anterior realice una validación del parámetro k en KNN, probando valores de k=1:10. Con el objetivo de obtener resultados consistentes, promedie los resultados de, al menos, 10 realizaciones correspondientes a diferentes particiones aleatorias de training/test. La representación de errores de train y test medios en función del parámetro k, puede ayudarle a seleccionar el valor de k óptimo en esta aplicación. Explique por qué el error de train es 0 cuando k=1. Incluya, además, en el documento a entregar, el código generado en esta parte de la práctica.

2.3.2 Método de Parzen

Con el objetivo de probar el método de ventanas de Parzen se facilita la función:

Predict_test = predict_parzen(X_train,Labels_train,N_classes,h,X_test)

Realice la siguiente prueba:

- Base de datos mediante transformación y N_dim seleccionados en el punto 4 del apartado anterior.
- Clasificador de Parzen con valores de parámetro h = 1, 10, 20 y 100.

Incluya en el documento a entregar el código generado en esta parte de la práctica y una tabla o una gráfica que contenga los errores obtenidos en train y en test, con el método Parzen para todos los valores del parámetro h propuestos. Comente los resultados obtenidos.

Como habrá observado, el entrenamiento con Parzen consume gran cantidad de tiempo. A pesar de ello se propone que si tiene tiempo pruebe algunos valores de sigma distintos a los propuestos y próximos al valor de sigma que haya producido el mínimo error de clasificación de test. Además, el promediado de resultados obtenidos mediante diferentes realizaciones (correspondientes a diferentes particiones aleatorias de training/test) le ayudará a obtener resultados más consistentes.

3 Base de Datos MicroArray

3.1 Características de la base de datos

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

```
NCI microarray data

Source and reference:

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

NCI microarray data

6830 genes
missing values have been imputed via SVD
60 cell lines, labels are below

1: CNS         5 samples
2: RENAL       7 samples
3: BREAST      9 samples
4: NSCLC       9 samples
5: UNKNOWN     1 samples
6: OVARIAN     6 samples
7: MELANOMA    8 samples
8: PROSTATE    2 samples
9: LEUKEMIA    6 samples
10:K562B-repro 1 samples
11:K562A-repro 1 samples
12:COLON       7 samples
13:MCF7A-repro 1 samples
14:MCF7D-repro 1 samples
```

La base de datos está formada por 64 vectores de dimensión d = 6830 muestras cada uno, presentando valores reales entre -10 y +10 aprox. Cada componente representa la presencia (+10) o ausencia (-10) de cada uno de 6830 genes distintos. Cada vector corresponde a una muestra de un órgano afectado por cáncer y proveniente de un paciente distinto. En función del tipo de cáncer, los vectores pertenecen a una de las 14 clases descritas. El nombre de MicroArray corresponde a la técnica utilizada para extraer cada una de las muestras.

El problema que existe con este tipo de bases de datos, radica principalmente en la enorme dimensionalidad de los vectores, que hace prohibitivo el uso de muchas de las técnicas de clasificación analizadas durante el curso (nótese que las matrices de covarianza estimadas serían forzosamente deficientes en rango (o singulares). Debido a este motivo, esta base de datos se analizará únicamente con el método K-Nearest.

3.2 Laboratorio

Ejecute el programa prac4_MA.m. En la BD que carga este fichero, se han seleccionado únicamente aquellas clases que tienen más de dos elementos por clase. El programa solicita el valor del parámetro *k_neig*=k, al principio de la ejecución.

Anote en una única tabla los errores obtenidos tanto para la base de train como para la base de test al variar: *k_neig* = 1,2,3,4. Razone los resultados obtenidos.

NOTA IMPORTANTE

Recuerde que en general las explicaciones deben apoyarse en los conceptos teóricos vistos en clase.