Miguel Ángel Navarro Arenas. 3CO11.

MEMORIA PRÁCTICAS PER - Parte 1

Ejercicios obligatorios.

En los ejercicios obligatorios se nos plantea completar el código de pca.m y los scripts pca+knn-exp.m y pca+knn-eva.m. Posteriormente, procedemos a ejecutar ambos scripts con un conjunto de datos de caracteres manuscritos, MNIST, los cuales se encuentran en cuatro archivos: train-images-idx3-ubyte.mat.gz, train-labels-idx1-ubyte.mat.gz, t10k-images-idx3-ubyte.mat.gz y t10k-labels-idx1-ubyte.mat.gz.

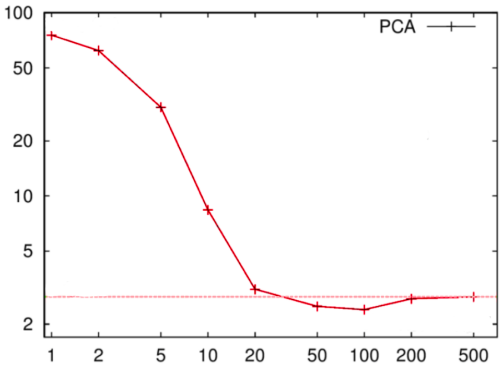
```
1-----------75.466667
2-----------62.150000
5-----------30.483333
10----------8.400000
20----------3.100000
50----------2.500000
100---------2.400000
200---------2.750000
500---------2.816667
```

Tras completar el código, realizamos una ejecución de ambos scripts, separando el conjunto de datos en entrenamiento (90%) y validación (10%). De ellos obtenemos los siguientes resultados:

```
Error PCA: 2.840000. Error no PCA: 3.090000
```

Primero ejecutamos el experiment. Podemos ver que el valor óptimo de k es 100, porque es con el que menos error se obtiene: 2,4. Con este valor de k, ejecutamos el pca+knn-eva.m.

Para ver cómo evoluciona el error, hemos realizado una gráfica con la distancia L2, obteniendo los siguientes resultados:



Vemos cómo desciende el error a medida que aumenta la k. Pero llega un punto que no mejora, sino que empeora. El punto óptimo es k = 100.

La línea horizontal representa el error mínimo. Por debajo de este error, significa que estamos redimensionando de tal manera que estamos quitando el ruido.

El hecho de que el error sea tan grande con unas k's pequeñas viene dada por la reducción de dimensionalidad de PCA y la posibilidad de que sobre una misma representación caigan dos datos que tienen una diferencia sustancial entre ellos.

Para comprobar nuestros datos, visitaremos la página oficial de MNIST, donde obtenemos los siguientes datos para nuestra prueba:

K-Nearest Neighbors			
K-nearest-neighbors, Euclidean (L2)	none	5.0	LeCun et al. 1998
K-nearest-neighbors, Euclidean (L2)	none	3.09	Kenneth Wilder, U. Chicago
K-nearest-neighbors, Euclidean (L2)	deskewing	2.4	LeCun et al. 1998
K-nearest-neighbors, Euclidean (L2)	deskewing, noise removal, blurring	1.80	Kenneth Wilder, U. Chicago

Vemos que el error de Kenneth Wilder es idéntico al que nosotros hemos obtenido en el test sin el PCA, 3.09. Vemos que nuestros valores, en este caso, son iguales que los de Kenneth y varían casi un 2% con respecto a los valores obtenidos por LeCun.

Posteriormente y con los dos siguientes datos de la tabla de la página de MNIST, podemos comprobar nuestros valores en las pruebas realizadas con PCA. El error que obtenemos se sale un poco de los parámetros que establece la página, pero solo dista un 0.4% del estudio de LeCun, por lo que nos podemos dar por satisfechos con estos resultados, aunque debemos saber que se puede mejorar incluso al 1.8% que vemos en los datos de Kenneth.