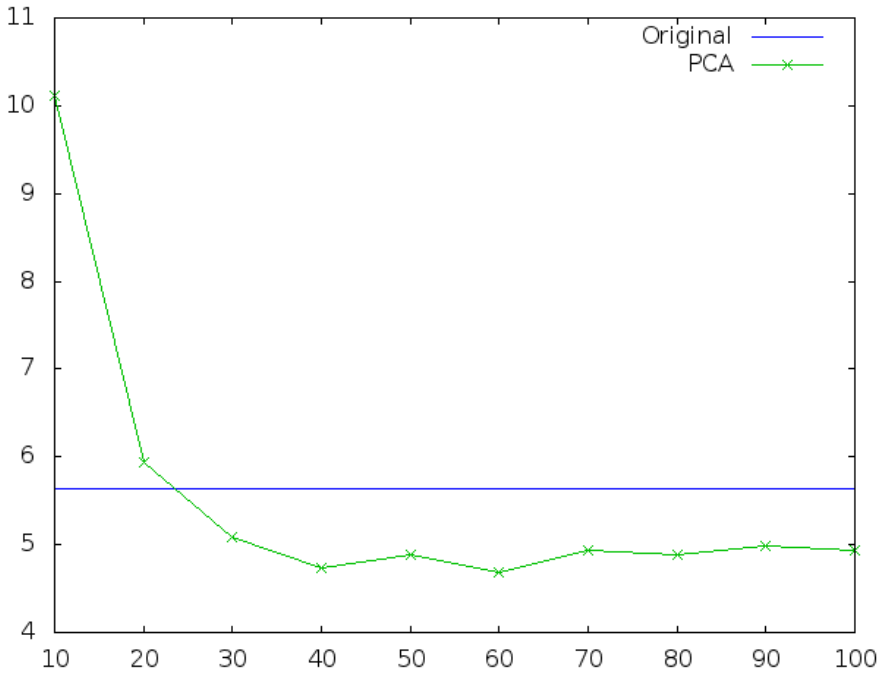
Practica 1. Aplicación de PCA y LDA a OCR
Josep Vicent Dols Dart

Tras haber haber implementado tanto PCA como LDA, y tras comprobar que las representaciones de sus 5 primeros vectores propios coinciden con las representaciones propuestas en ambos casos, se pasa a comprobar la tasa de error de clasificación que proporcionan para el conjunto de datos propuesto. Para obtener dicho error usamos un clasificador por vecinos más cercanos, implementado en la función knn incluida en el material.

El conjunto en su forma original, sin aplicar ningún cambio sobre él, tiene un error de clasificación del 5,63%. El problema de dicho conjunto es su alta dimensionalidad(256), por lo que es recomendable usar diferentes técnicas para reducirla.

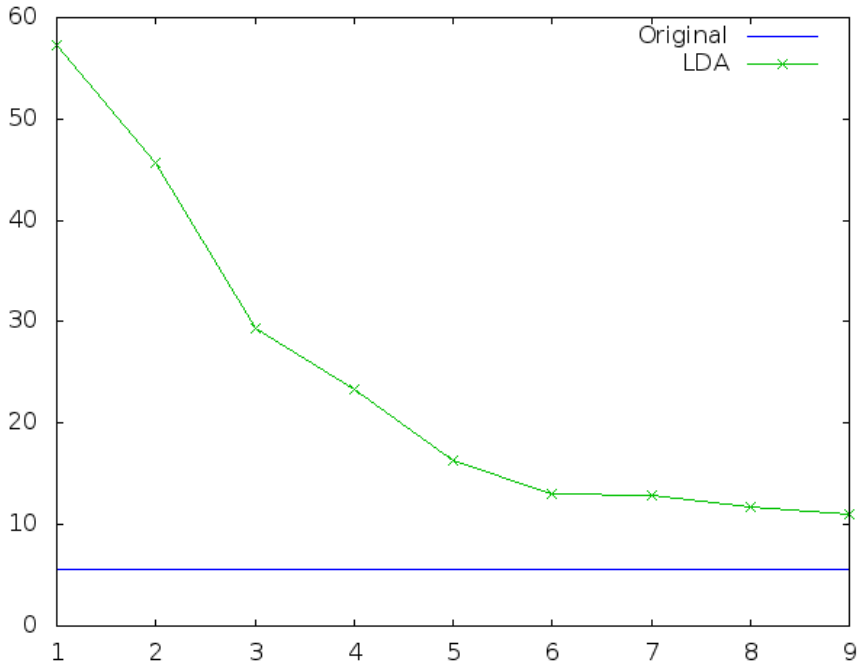
La primera propuesta es usar PCA; para comprobar el efecto que tiene la reducción de dimensiones en el error, se usa el script pcaexp.m, dentro del cual esta implementado un fragmento de código el cual calcula el error de clasificación de un rango de nuevas dimensiones para los datos, tras haber aplicado PCA. En esta caso, se calculan las dimensiones de 10 a 100, con saltos de 10 en 10. Tras la ejecución, se obtienen estos resultados:



Se puede observar como para los casos en el que la reducción de dimensiones es considerable, el error aumenta, sin embargo, para dimensiones superiores a 30, se puede llegar a la conclusión de que la reducción de dimensionalidad disminuye el posible error de clasificación.

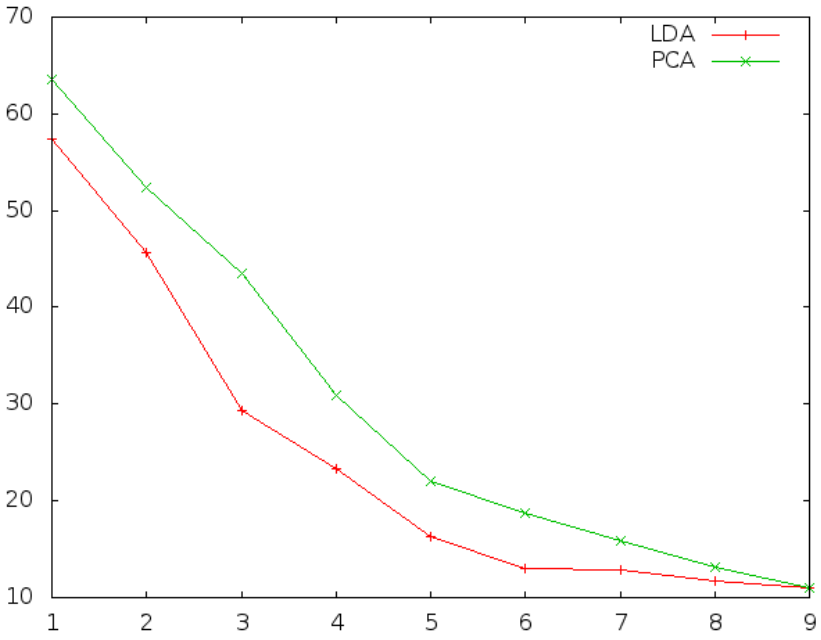
La segunda propuesta es usar LDA; al igual que en el caso anterior, se ha reutilizado el script pcaexp.m, usando esta vez el fragmento de código que se encuentra comentado, correspondiente a LDA. Al usar LDA, el numero de dimensiones a las que se reduce el

conjunto original debe de ser un valor entre 1 y el numero de clases menos 1, que en este caso es 9. Para este caso, se calculan todas las dimensiones que hay en dicho rango de valores, obteniendo el siguiente resultado:



En este caso, se puede observar como el error es bastante superior al original, especialmente para dimensiones menores que 6. Esto es debido a que en este conjunto solo tiene 10 clases, por lo que una reducción bastante amplia provoca que algunas clases dejen de ser separables entre ellas, aumentando el error de clasificación.

Sin embargo, como se puede ver en la ultima grafica, si se aplica la reducción mediante PCA para este mismo rango de valores, se puede comprobar que el error es superior para todos los casos respecto al obtenido con LDA.



Con los resultados de la ultima grafica, se puede llegar a la conclusión de que en el caso de reducir a una dimensión mayor al numero de clases, la única opción es PCA, pero si se la reducción es a una dimensión menos, pese a que es posible con ambas, la mejor opción es con LDA. Este hecho es el que se usa para más tarde usar ambas reducciones a la vez, reduciendo primero con PCA al valor más conveniente, y a ese nuevo conjunto de datos con una dimensión reducida, aplicarle LDA para volver a reducir la dimensión, esta vez a un valor menor al numero de clases.