Curso 2021-2022

Iñaki Diez Lambies y Manuel Diaz Pastor

Percepción

Proyecto de prácticas

Reconocimiento de dígitos manuscritos: MNIST

Contenido

[2 Ejercicio 2.1 - Principal Component Analysis 2](#_Toc101820031)

[2.1 Ejercicio 2.2 – Comprobación PCA 2](#_Toc101820032)

[3 Ejercicio obligatorio – KNN + PCA 2](#_Toc101820033)

[3.1 pca+knn-exp.py 2](#_Toc101820034)

[3.1.1 Resultados 3](#_Toc101820035)

[3.2 pca+knn-eva.py 3](#_Toc101820036)

[3.2.1 Resultados 3](#_Toc101820037)

[4 Ejercicio opcional – Algoritmo de Wilson 4](#_Toc101820038)

[4.1 mnn 4](#_Toc101820039)

[4.2 knnV 4](#_Toc101820040)

[4.3 wilson 4](#_Toc101820041)

[4.4 pca+knn+wilson-exp.py 5](#_Toc101820042)

[4.4.1 Resultados 5](#_Toc101820043)

[4.5 pca+knn+wilson-eva.py 5](#_Toc101820044)

[4.5.1 Resultados 5](#_Toc101820045)

Entrega 2

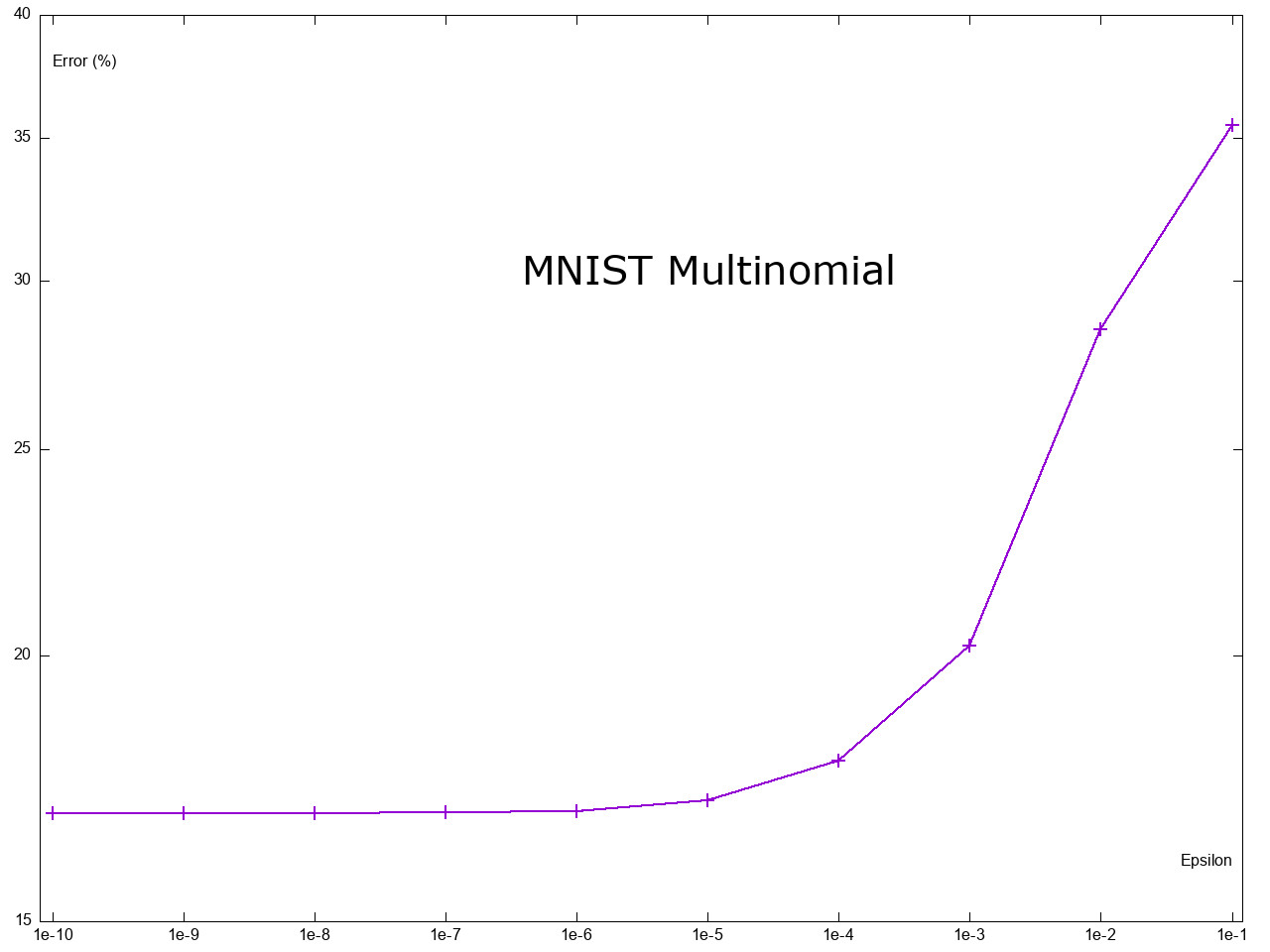
# Ejercicio obligatorio: Clasificador multinomial

Para la realización de esta entrega se nos ha pedido preparar un algoritmo de clasificación multinomial que recibe como argumento un conjunto de datos y etiquetas de entrenamiento, un conjunto de datos y etiquetas de validación y un vector con una serie de valores de suavizado para el algoritmo de suavizado de Laplace.

Para el proceso de la función multinomial de multinomial.py hemos dividido el procedimiento en dos partes: cálculo de parámetros y clasificación con suavizado.

## Multinomial-exp.py

Además de esto, hemos realizado un experimento en *multinomial-exp.py* para comprobar el impacto de los diferentes valores de suavizado, siendo estos 1e-10, 1e-9, 1e-8 1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2 y 1e-1, en la clasificación del 10 % de los datos que hemos usado para test. A la vista de los resultados hemos obtenido la siguiente gráfica;



Como podemos observar, el resultado nos indica que a mayor valor de suavizado peores porcentajes de error genera. Esto se debe principalmente a que la mayor funcionalidad del suavizado en este caso es evitar fallos en la ejecución por la abundancia de ceros en la matriz.

Por lo que, en conclusión, extraemos que el mejor valor de suavizado para este conjunto de datos con la distribución multinomial es cualquiera entre 1e-10 y 1e-8.

## Multinomial-eva.py

Una vez obtenido el valor de suavizado 1e-10, hemos realizado la evaluación del clasificador para el conjunto de datos de validación con este número mediante el código *multinomial-eva.py*. El porcentaje de error en la clasificación del clasificador multinomial para el conjunto de datos MNIST es del 16’32%.

Si comparamos el resultado obtenido con los documentados en la web MNIST podremos comprobar que resulta ser el peor analizado, siendo superado incluso por los tres clasificadores lineales registrados (por pares y de una capa). También podemos concluir que resulta una pérdida de eficacia considerable con respecto a los clasificadores de vecinos más cercanos vistos en la primera parte de este proyecto.

Tabla

Descripción generada automáticamente

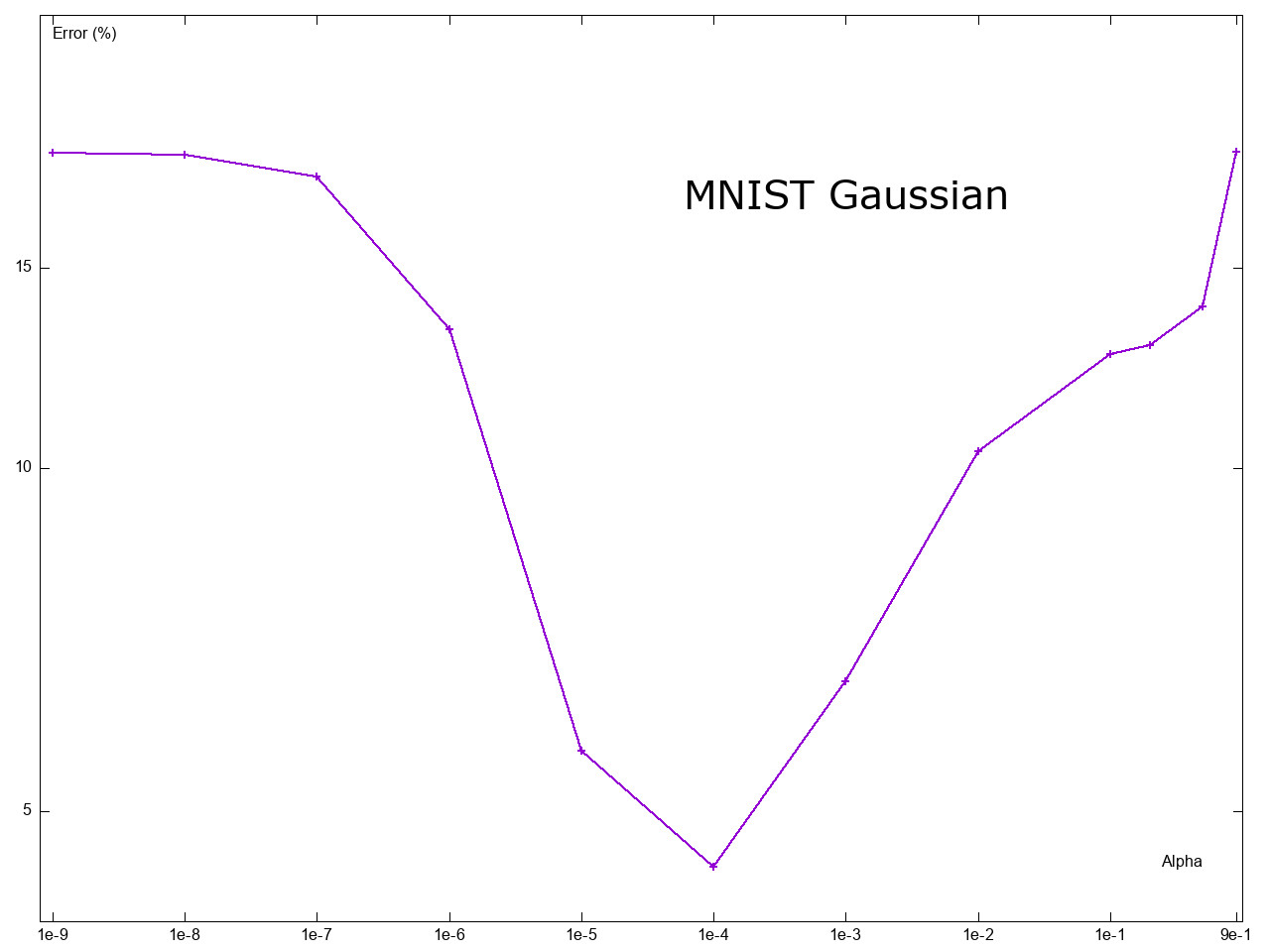
# Clasificador Gaussiano

Para la segunda pregunta de esta segunda parte del proyecto, hemos construido un clasificador gaussiano en gaussian.py que toma como argumentos le conjunto de entrenamiento, sus etiquetas, el conjunto de prueba, sus etiquetas y un grupo de valores de alfa como valores posibles de suavizado para utilizar el algoritmo de “flat-smoothing”.

## Gaussian-exp.py

Una vez resueltos todos los problemas encontrados en el desarrollo del código, hemos preparado un experimento en *gaussian-exp.py*, al igual que con el clasificador multinomial donde pasamos los diferentes valores de suavizado 1e-9, 1e-8 1e-7, 1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1 y 9e-1 para la clasificación del 10% de los datos de entrenamiento de MNIS a partir del 90% restante.

El resultado del experimento se expone en la siguiente gráfica, donde el eje “x” representa los valores de suavizado probados y el eje “y” el porcentaje de error resultante.



A la vista de los resultados, podemos comprobar que el valor de suavizado óptimo para la clasificación de estos datos es 1e-4.

## Gaussian-eva.py

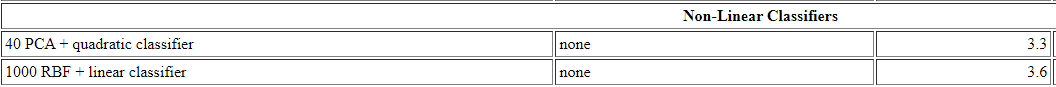
Una vez obtenido el valor óptimo de suavizado, hemos de realizar la evaluación mediante el código *gaussian-eva.py* que se encarga de probar la clasificación del conjunto de datos de prueba de MNIST en base al conjunto de entrenamiento para un valor de alfa concreto.

Como hemos mencionado previamente, el valor seleccionado es 1e-4. Teniendo esto en cuenta, el porcentaje de error en la evaluación del clasificador para MNIST mediante un clasificador gaussiano es del 4’18%. Llama la atención la mejora significativa de este clasificador cuadrático con respecto al clasificador lineal. Además, si volvemos a comprobar los resultados con los existentes en la base de datos de MNIST, podremos observar que si bien la mayoría de ellos siguen superando al clasificador gaussiano, obtenemos mejores resultados que para los clasificadores lineales y el clasificador de k-vecinos más cercanos básico con distancia L2:

Tabla

Descripción generada automáticamente

Sin embargo, aún queda por detrás de los clasificadores no lineales que aparecen en la página:



# Ejercicio Opcional: PCA+Gaussian.py

Tras obtener los resultados de los mejores valores de suavizado de alfa para el suavizado “flat smoothing” con los datos en bruto, hemos de realizar un análisis de este mismo proceso para los datos reducidos mediante la técnica de Principal Component Analysis desarrollado previamente. Es de esta forma que hemos probado todos los valores de dimensiones para PCA probados para el clasificador de vecinos más cercanos.

## PCA+Gaussian-exp.py

De esta forma obtenemos los resultados en una gráfica donde cada curva representa los valores de PCA utilizados, el eje “x” el valor de suavizado empleado y el eje “y” porcentaje de error resultante.

El código genera la matriz de proyección de datos para cada valor de “k” (dimensiones en PCA) y llama a la función de clasificador gaussiano con los datos proyectados y el vector de valores de suavizado.

De esta forma obtenemos los siguientes resultados para todas las combinaciones contempladas:

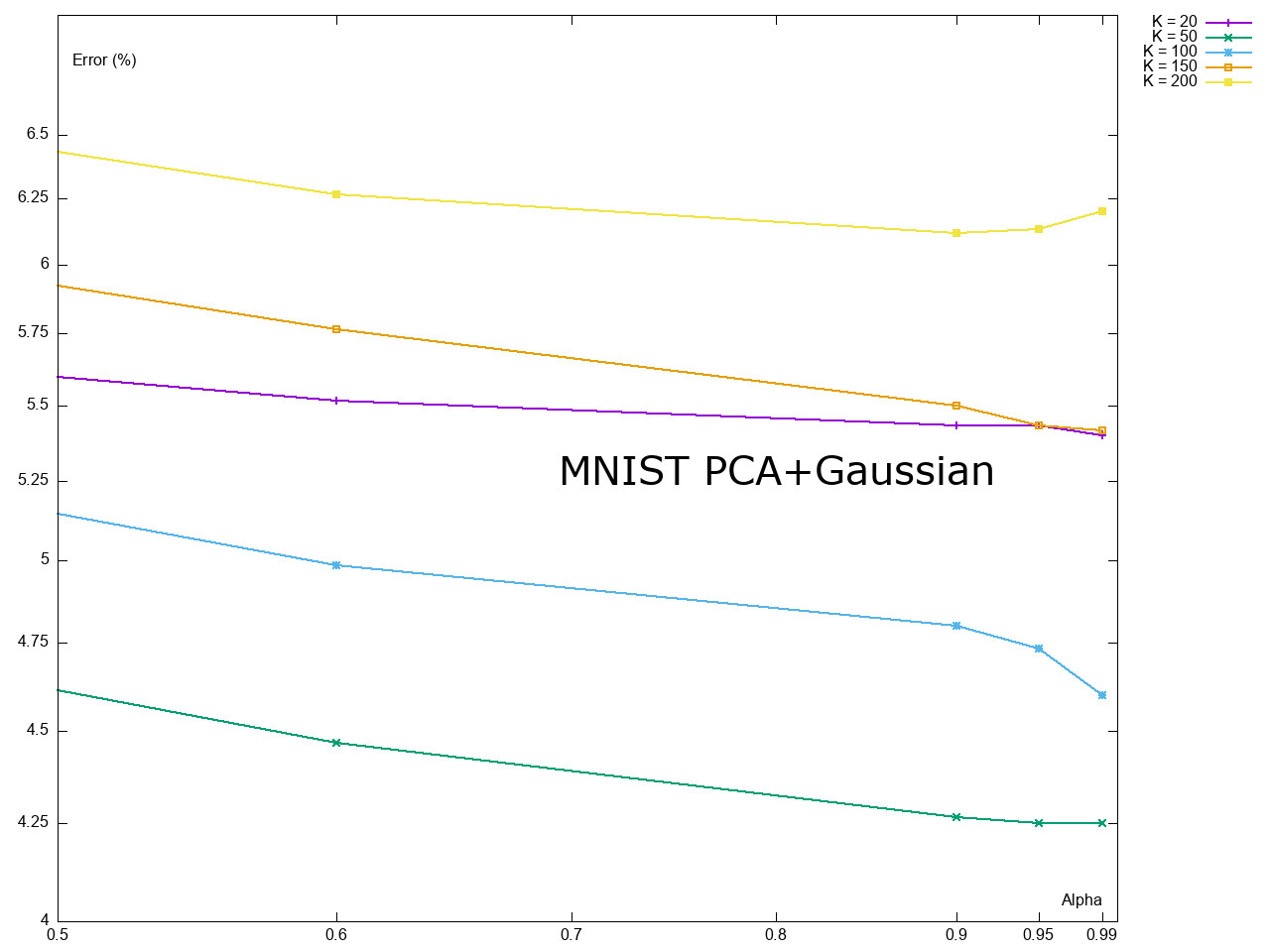
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| alphas\Ks | K=1 | K=2 | K=5 | K=10 | K=20 | K=50 | K=100 | K=150 | K=200 | K=500 |
| 1,00E-09 | 70,65 | 58,08 | 37,12 | 25,30 | 20,35 | 19,10 | 18,93 | 18,95 | 18,95 | 18,93 |
| 1,00E-08 | 70,67 | 58,08 | 37,08 | 25,30 | 20,32 | 19,07 | 18,88 | 18,92 | 18,90 | 18,85 |
| 1,00E-07 | 70,68 | 58,08 | 36,97 | 24,87 | 19,88 | 18,38 | 18,18 | 18,08 | 18,03 | 18,03 |
| 1,00E-06 | 70,62 | 57,83 | 35,20 | 22,30 | 15,83 | 13,68 | 13,38 | 13,30 | 13,27 | 13,25 |
| 1,00E-05 | 70,55 | 56,37 | 29,60 | 14,67 | 8,20 | 5,88 | 5,62 | 5,62 | 5,62 | 5,65 |
| 0,0001 | 70,37 | 56,08 | 27,60 | 11,70 | 5,62 | 4,43 | 4,40 | 4,40 | 4,42 | 4,47 |
| 0,001 | 70,27 | 56,25 | 27,83 | 12,00 | 6,22 | 5,40 | 5,40 | 5,80 | 6,03 | 6,48 |
| 0,01 | 70,03 | 56,20 | 27,87 | 12,10 | 6,57 | 6,10 | 6,87 | 7,70 | 8,15 | 10,12 |
| 0,1 | 69,97 | 55,88 | 27,80 | 11,92 | 6,28 | 5,77 | 6,53 | 7,40 | 8,15 | 11,82 |
| 0,2 | 70,20 | 55,58 | 27,68 | 11,75 | 6,03 | 5,43 | 6,05 | 6,78 | 7,35 | 11,77 |
| 0,6 | 69,92 | 54,88 | 27,22 | 11,25 | 5,52 | 4,47 | 4,98 | 5,77 | 6,27 | 12,22 |
| 0,9 | 70,05 | 54,90 | 27,50 | 11,22 | 5,43 | 4,27 | 4,80 | 5,50 | 6,12 | 13,50 |
| 0,95 | 70,08 | 54,85 | 27,52 | 11,20 | 5,43 | 4,25 | 4,73 | 5,43 | 6,13 | 13,85 |
| 0,99 | 70,02 | 54,95 | 27,52 | 11,22 | 5,40 | 4,25 | 4,60 | 5,42 | 6,20 | 14,70 |

Como podemos observar en la tabla, el mejor resultado obtenido para el clasificador nos indica que la mejor combinación para el clasificador gaussiano es con suavizado con alfa 0’95 y 50 dimensiones de PCA. Esto resulta extremadamente interesante, puesto que un alto valor de alfa para el suavizado en el clasificador cuadrático sin PCA significaba un mayor porcentaje de error. Sin embargo, como se puede observar también en las siguientes tablas, los resultados obtenidos indican lo contrario con la disminución de dimensionalidad. Es decir, al contrario que ocurría previamente, el mejor resultado se obtiene con un alto valor de alfa. Esto probablemente se debe a la influencia de la reducción de datos para mantener la máxima dispersión en los datos.

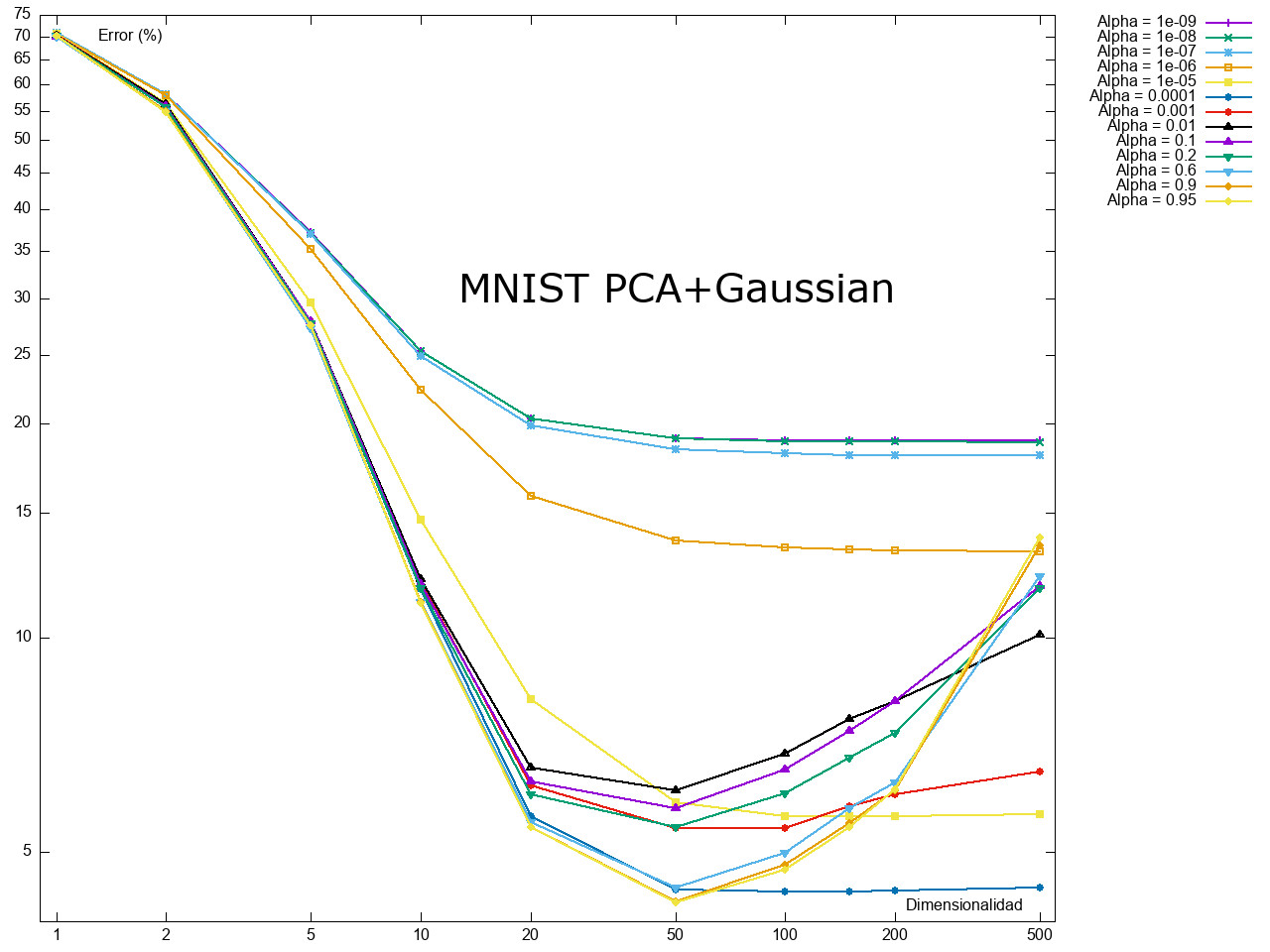
Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfica con todas las combinaciones.

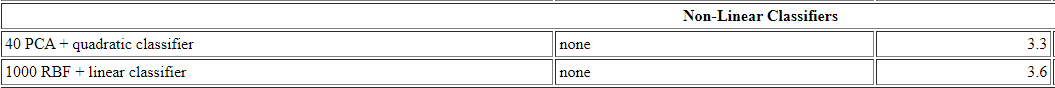


Gráfica centrada en los valores más relevantes.



Gráfica respecto al suavizado

De esta forma, con el valor de alfa 0’95 y 50 dimensiones, nos disponemos a preparar la evaluación de los datos de prueba de MNIST. El resultado obtenido nos indica que el porcentaje de error para un clasificador gaussiano es del 3’7%. Esto implica una mejoría con respecto a la versión sin reducción de dimensionalidad, acercando más este resultado a los obtenidos por clasificadores cuadráticos registrados en la página de MNIST:



Siendo estos muy próximos al segundo clasificador y algo menos al cuadrático con 40 dimensiones por PCA.