



DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Métodos Numéricos

Trabajo Práctico 2

Reconocimiento de dígitos.

Resumen

En este trabajo pondremos en práctica distintos algoritmos para reconocer una cierta cantidad de dígitos manuscritos. Se trabajará con una base train, a la cual le realizaremos particiones para entrenar los algoritmos. Se implementará el método de kNN (k -Nearest Neighbors). Debido a que este es sensible a la dimensión de los objetos a considerar, además se implementará el método de Análisis de Componentes Principales para reducir el tamaño de dichos objetos. Se llevarán a cabo experimentaciones para poder determinar los parámetros óptimos para cada método. Al final del trabajo, se llegarán a conclusiones sobre lo descubierto.

Integrante	LU	Correo electrónico
Armagno, Julián	377/12	julian.armagno@gmail.com
More, Ángel	931/12	angel_21_fer@hotmail.com
Pinzón, Germán	475/13	pinzon.german.94@gmail.com
Porto, Jorge	376/11	cuanto.p.p@gmail.com

Palabras claves:

Learning Machine. kNN . Análisis de Componentes Principales. Auto-Valores. Auto-Vectores. Método de las Potencias. K-fold cross validation

Índice

1. Introducción Teórica	3
1.1. k Nearest Neighbor y Principal Component Analysis	3
1.2. Método de la potencia y deflación	4
2. Desarrollo	5
2.1. k-Nearest Neighbors (kNN):	5
2.2. Principal Component Analysis (PCA):	6
2.3. K-Fold Cross Validation	8
2.4. Evaluación de los métodos	10
2.5. Elección de K, hipótesis 1	10
2.6. Experimento 1: Variando el k en kNN	11
2.7. Experimento 2: PCA + KNN	11
2.7.1. Variando k	11
2.7.2. Variando α	11
3. Resultados	13
3.1. Experimento 1	13
3.2. Experimento 2	15
3.2.1. Variando k	15
3.2.2. Variando α	17
4. Discusión	20
4.1. Hipótesis 1:	20
4.2. Experimento 1:	21
4.3. Experimento 2:	21
4.3.1. Variando k	21
4.3.2. Variando α	21
4.4. Competencia Kaggle	22
5. Conclusiones	23
6. Apéndices	24
6.1. Apéndice B:	28

1. Introducción Teórica

1.1. k Nearest Neighbor y Principal Component Analysis

En el presente trabajo utilizaremos los métodos de kNN y PCA para llevar a cabo el reconocimiento de dígitos manuscritos. Comenzaremos explicando resumidamente los métodos utilizados, de una manera más general. Dado un vector v y un conjunto de vectores U , queremos saber a que clase pertenece v (qué dígito es). Nosotros sabemos a que clase pertenece cada vector $u_i \in U$, entonces una manera sencilla de "estimar" a que clase pertenece v en base a la información que poseemos, es comparándolo con cada vector de U . Una manera de efectuar esta comparación es tomando la distancia (norma 2) entre v y cada vector $u_i \in U$ y quedarnos con la clase del u_m que minimice esa distancia sobre todos los demás. Lo que hace el algoritmo kNN es generalizar un poco esta idea. En lugar de quedarnos con la clase del u_m que minimice la distancia a v , kNN toma las clases de los k vectores u_1, u_2, \dots, u_k cuyas distancias sean mínimas y elige entre ellas la clase que más aparezca. Para conocer el parámetro k es necesario realizar experimentos probando distintos valores y tomar una decisión en base a la calidad de los resultados.

El algoritmo kNN es conceptualmente fácil de entender e implementar, lo cual constituye una ventaja, pero su desventaja principal es el tiempo de cómputo que requiere. Esto se debe a que, como los vectores representan imágenes, la dimensión de estos vectores suele ser grande. Dado que la idea detrás del reconocimiento de imágenes en este trabajo radica en utilizar una base de datos relativamente grande (de ahora en más dbTrain) para determinar que tipo de imagen es la que estamos tratando de reconocer, el costo de computar estas distancias se multiplica por la cantidad de imágenes que tenemos en nuestra base de datos. Vemos que este método, así como está planteado, no escala.

Como ya dijimos, por cuestiones de performance no es conveniente utilizar kNN de manera directa con los datos que tenemos. Lo que vamos a intentar hacer entonces, es realizar una transformación de nuestros datos de manera tal que luego, cuando queramos aplicar kNN no nos resulte tan costoso, este método es conocido como PCA (Principal Component Analysis). Recordemos que nuestros datos son vectores en un espacio \mathbb{R}^n , entonces lo que vamos a querer hacer es transformarlos a vectores en otro espacio \mathbb{R}^α tal que $\alpha < n$. Pero también vamos a querer que esta transformación no "cambie por completo" a nuestros vectores, en el sentido de que sigan representando las imágenes que representaban o al menos conserven las "partes relevantes" de ellas. Para poder llevar a cabo esta transformación realizaremos los siguientes pasos:

1. Tomamos los vectores de dbTrain que usaremos para comparar la imagen que queremos reconocer en forma matricial (cada vector una fila de la matriz) y hallar la matriz de covarianza $M \in \mathbb{R}^{n \times n}$.
2. Hallar una matriz $P \in \mathbb{R}^{n \times n}$ que nos permita disminuir la covarianza de los datos de dbTrain. Esto equivale a buscar las variables que tengan la mayor varianza entre sí y covarianza 0. El objetivo de esto es disminuir la redundancia en nuestros datos.
3. Sea P' la matriz P con las primeras α columnas (este parámetro se fija mediante experimentación), es decir $P' \in \mathbb{R}^{n \times \alpha}$, y x_i la i -ésima imagen de dbTrain, realizamos el producto $x'_i = P'^t \hat{x}_i$, ahora x'_i es nuestra nueva i -ésima imagen de train y está en el espacio $\mathbb{R}^{n \times \alpha}$.
4. Sea x una imagen vectorizada cuya clase queremos reconocer, realizamos el producto $x' = P'^t \hat{x}$ donde $\hat{x} = (x - \mu) / (\sqrt{n - 1})$, μ es la media de las imágenes de dbTrain y n la cantidad de imágenes de dbTrain. Como ahora $\hat{x} \in \mathbb{R}^\alpha$ y los vectores de dbTrain están en el mismo espacio, podemos aplicar kNN con x' y los x'_i .

Para obtener P lo que hacemos es hallar la base ortonormal de autovectores de M . Sabemos que dicha base existe por ser M simétrica. Entonces la columna i de P va a ser el vector v_i , el i -ésimo autovector de M . Esta base lo que nos permite hacer es "observar" a nuestros datos desde otro lugar, o sea ahora nuestros ejes de coordenadas van a estar en las direcciones donde más varianza existe entre los datos.

Una vez completados estos tres pasos, cuando queramos reconocer a que clase pertenece un vector v , trabajamos con $v' = P^t v$ y, dado que ahora v' es un vector de \mathbb{R}^α al igual que los vectores de dbTrain podemos aplicar kNN .

1.2. Método de la potencia y deflación

En la sección anterior explicamos los métodos que nos permiten reducir la dimensión de las imágenes con las cuales vamos a trabajar y, en base a cierta información que tenemos, efectuar comparaciones para predecir a que clase pertenece una imagen. Vimos que, uno de los pasos de PCA es obtener una matriz P cuyas columnas son los autovectores de otra matriz M . El método de la potencia, junto con el método de deflación, nos permite encontrar los autovectores y autovalores asociados a la matriz M .

Dada una matriz $A \in \mathbb{R}^{n \times n}$ cuyos autovalores $\lambda_1, \dots, \lambda_n$ cumplen $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|$ y v_1, \dots, v_n los autovectores asociados, el Método de la potencia estimará v_1 y λ_1 . Decimos estimará porque para tener el valor exacto deberíamos calcular un límite, entonces vamos a "simular" este límite computacionalmente con lo cual el resultado puede tener cierto error. Este método necesita de un vector inicial x_0 y un valor de k relativamente grande y lo que va a hacer es calcular $x_{i+1} = \frac{Ax_i}{\|Ax_i\|_2}$ y $\hat{\lambda}_1 = \frac{\Phi(Ax_{i+1})}{\Phi(Ax_i)}$ desde $i = 1$ hasta k . El resultado será $x_{k+1} = \hat{v}_1 \approx v_1$ y $\hat{\lambda}_1 \approx \lambda_1$. Para que esto se cumpla es importante aclarar que la función $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$ que usamos debe cumplir las siguientes propiedades:

- Debe ser continua
- Si $v \in \mathbb{R}^n$ y $v \neq 0$ entonces $\Phi(v) \neq 0$
- Si $v \in \mathbb{R}^n$ y $\alpha \in \mathbb{R}$ entonces $\Phi(\alpha v) = \alpha \Phi(v)$

El Método de la potencia también pide que el vector inicial x_0 no sea perpendicular a v_1 , sin embargo a la hora de implementarlo en una computadora esto puede no ser tenido en cuenta y no traerá grandes consecuencias. Esto se debe a que nuestro resultado va a estar arrastrando cierto error a medida que el método itere y este error va a estar cambiando la dirección de x_i (aunque sea mínimamente) y en ese caso dejaría de ser perpendicular a v_1 .

Vimos como funciona el Método de la potencia y que nos devuelve el autovalor de mayor módulo junto con su autovector asociado, sin embargo nosotros queremos obtener todos los autovalores y autovectores de A . Esto se soluciona definiendo una nueva matriz \hat{A} y aplicando nuevamente el Método de la potencia pero ahora sobre \hat{A} . Supongamos que ya tenemos los valores $\hat{v}_1 \approx v_1$ y $\hat{\lambda}_1 \approx \lambda_1$ obtenidos al aplicar el Método de la potencia sobre A , entonces para obtener \hat{v}_2 y $\hat{\lambda}_2$ aplicamos nuevamente el método pero ahora sobre $\hat{A} = A - \hat{\lambda}_1 \hat{v}_1 \hat{v}_1^t$. Para el caso general, si queremos obtener \hat{v}_{i+1} y $\hat{\lambda}_{i+1}$, tenemos \hat{v}_i y $\hat{\lambda}_i$ y nuestra matriz es \hat{A} , definimos $\hat{\hat{A}} = \hat{A} - \hat{\lambda}_i \hat{v}_i \hat{v}_i^t$ y aplicamos el método sobre $\hat{\hat{A}}$. Esto se le llama *deflación*. Los autovalores y autovectores de la nueva matriz que definimos van a ser los mismos que la matriz original (la primera o la que definimos en el paso anterior) excepto por el autovalor de mayor módulo y su autovector asociado.

Si bien vimos que para aplicar el método de la potencia en una matriz A se tiene que cumplir que A tenga un autovalor mayor estricto (de multiplicidad 1) en módulo a todos los demás, para poder hacer deflación iterativamente y obtener todos los autovalores y autovectores de A es necesario que valgan las desigualdades de manera estricta: $|\lambda_1| > |\lambda_2| > |\lambda_3| > \dots > |\lambda_n|$. Esto debe ser así para que cada nueva matriz que definimos cumpla las hipótesis que requiere el Método de la potencia.

2. Desarrollo

2.1. k-Nearest Neighbors (kNN):

El primer método utilizado es kNN. Como se dijo en la introducción teórica, este método es muy intuitivo, se basa en tomar a cada imagen como un punto en el espacio, y se hace una votación de la moda de los k vecinos cercanos. Todas las imágenes de dbTrain (28x28 píxeles), se encuentran guardadas vectorizadas en una matriz de $\mathbb{R}^{42000 \times 785}$. La primera columna establece el label de cada imagen, las columnas restantes son los píxeles.

Mediante el K que determine la partición, utilizando la técnica de K-Fold Cross Validation, vamos a dividir esa matriz, en dos, una de train, con $(42000/K) * (K - 1)$ imágenes y otra de test, con $42000/K$ imágenes. A su vez, vamos a eliminar la primera columna de cada matriz, y nos las vamos a guardar en 2 vectores diferentes. Así podremos comparar cada imagen de la matriz de test, con todas las imágenes de la matriz de train y al mismo tiempo tener guardados los labels correspondientes a cada imagen para luego comparar si cada imagen de test fue caracterizada correctamente. Finalmente la idea algorítmica se detalla en el siguiente pseudocódigo:

Algoritmo 1 kNN(int k , matriz Test, matriz Train, vector digTest, vector digTrain)

```

1: reconocidos = 0
2: for z = 0; z < Test.filas() ; z++ do
3:   vector<pair<int, int> > normas2AlCuadrado
4:   \\ La primera componente de cada tupla es el label del dígito de la base de train (de ahora en más d), y la segunda componente es la distancia entre el dígito de la base de test y d
5:   for m = 0; m < Train.filas(); m++ do
6:     \\ calculamos las distancias
7:     distanciaAlCuadrado = 0
8:     for i = 0; i < Test.columnas() ; i++ do
9:       distanciaAlCuadrado += (Test[z][i] - Train[m][i]) *(Test[z][i] - Train[m][i])
10:    end for
11:    if normas2AlCuadrado.size() < k then
12:      \\ colocamos las primeras k normas
13:      pair<unsigned int, int> a
14:      a.first = digTrain[m]
15:      a.second = distanciaAlCuadrado
16:      normas2AlCuadrado.pushback(a)
17:    else
18:      if hayMayor(normas2AlCuadrado, distanciaAlCuadrado) then
19:        \\ si ya tengo k voy sacando las mayores distancias
20:        int posMayor = dondeMayor(normas2AlCuadrado)
21:        normas2AlCuadrado[posMayor].first = digTrain[m]
22:        normas2AlCuadrado[posMayor].second = distanciaAlCuadrado
23:      end if
24:    end if
25:  end for
26: end for
27: \\ Ahora hago la votación entre los k vecinos
28: digitos[10] = 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
29: for int t = 0; t < k; t++ do
30:   digitos[normas2AlCuadrado[t].first]++
31: end for
32: ganador = 0
33: for int x = 0; x < 10; x++ do
34:   if digitos[x] > digitos[ganador] then
35:     ganador = x
36:   end if
37: end for
38: \\ Si en la votación gana el verdadero label de la imagen de test, sumamos reconocidos
39: if ganador == digTest[z] then
40:   reconocidos++
41: end if
42: tasaDeReconocimiento = reconocidos/Test.filas()

```

2.2. Principal Component Analysis (PCA):

El segundo método utilizado es PCA. Dado que las imágenes están representadas por vectores y las dimensiones de los mismos pueden ser muy grandes, a la hora de analizar las imágenes se requiere una gran cantidad de tiempo de cómputo así como de memoria. PCA tiene como objetivo redimensionar dichos vectores, siempre que se obtenga un determinado compromiso entre la nueva cantidad de variables y la calidad de las imágenes que ahora se representan. Para esto se construye un nuevo sistema de ecuaciones donde en el eje i se representa la i -ésima varianza de mayor valor para el conjunto original de imágenes dadas. De esta manera cada variable está correlacionada,

siendo las primeras las de mayor importancia, por lo que se pueden obviar las últimas componentes (ya que serían las que menos importancia tienen) reduciendo el número de variables. Para poder llevar a cabo esta transformación lineal es necesario construir la matriz de covarianza para el conjunto de valores que representan a las imágenes. Para esto, sean n observaciones (imágenes) y dadas dos coordenadas x_i, x_k su covarianza se obtiene como:

$$\sigma_{x_j x_k} = (1/(n-1)) \sum_{i=1}^n (x_j^{(i)} - \mu_j)(x_k^{(i)} - \mu_k) = (1/(n-1))(x_k^{(i)} - \mu_k v)^t (x_j^{(i)} - \mu_j v)$$

con μ_j, μ_k la media de las coordenadas j y k respectivamente; $x_j^{(i)}, x_k^{(i)}$ las coordenadas j y k de la i -ésima imagen vectorizada y $v^t = (1, \dots, 1)$.

De esta forma la covarianza entre dos muestras puede ser calculado como el producto entre dos vectores. Por lo que si $A \in \mathbb{R}^{n \times 784}$ representa n imágenes de train vectorizadas de 784 variables cada una, la covarianza pueden ser calculadas como:

$$M_x = A^t * A'; \quad \text{con } A' = A/\sqrt{n-1} \text{ y } M_x \in \mathbb{R}^{784 \times 784}$$

obteniendo en la diagonal la varianza de cada coordenada. Con el fin de disminuir las redundancias (covarianza), se plantea un cambio de base. Para esto se diagonaliza la matriz M_x , de forma tal que:

$$M_x = P * D * P^t; \quad D \text{ diagonal y } P \text{ matriz ortogonal con los autovectores de } A \text{ como columnas}$$

Para obtener los autovectores de M_x se empleó el método de la potencia. Dado que la misma es una técnica iterativa que converge al autovector asociado, se estableció que la misma se detenga luego de alcanzar un máximo de 3000 iteraciones o cuando la diferencia entre cada coordenada del vector obtenido en la iteración k respecto de la $k+1$ sea muy poca (en este caso optamos por 0.0000009). Una vez obtenido el primer autovector procedimos a calcular el segundo, aplicando el método de deflación. Ambas técnicas serán aplicadas un número α de veces. En la sección experimentación se evaluará como se comportan los resultados y el tiempo al variar este valor. Una vez obtenidos los α autovectores se procedió a llevar a cabo el cambio de base, tanto para las imágenes utilizadas como train como para las de test, por lo que se multiplicó cada autovector obtenido por las imágenes vectorizadas en A' :

$$v_i^t * a^{(i)}; \quad v_i \text{ el } i\text{-ésimo autovector y } a^{(i)} \text{ la } i\text{-ésima imagen vectorizada}$$

Lo que matricialmente se puede obtener al hacer: $A' * P$ (llamaremos a la matriz resultante $TcTrain$). De esta forma la matriz $TcTrain \in \mathbb{R}^{n * \alpha}$. Para poder comparar las imágenes de test (matriz B) con $TcTrain$, es necesario llevar a las imágenes de test a las mismas dimensiones, por lo que a cada una se le restará la media que corresponda y dividirá por $\sqrt{n-1}$, obteniendo B' . Una vez realizado esto procedemos a cambiar la base tal como se hizo para las de train: $TcTest = B' * P$. En las matrices $TcTrain$ y $TcTest$ se encuentran las nuevas coordenadas de las imágenes (con reducción de la dimensión) almacenadas como filas.

Finalizado los cambios de bases obtenemos imágenes en un mismo espacio vectorial cuyas dimensiones son menores a las originales. Pudiendo aplicar distintas técnicas para llevar a cabo el reconocimiento de los dígitos. Optaremos por aplicar KNN para así poder comparar las diferencias entre tiempo de cómputo y análisis cuando se busca reconocer dígitos utilizando PCA + KNN y solo KNN.

A continuación se muestra un pseudocódigo de los procedimientos descriptos anteriormente:

Algoritmo 2 PCA(matriz Train, matriz Test, int α)

```

1: n = cantidad de imágenes de train
2: \\ Se procede a calcular el promedio de las imagenes de train
3: vectorPromedio  $\leftarrow$  crear(784)
4: for i = 0; i < 784; i++ do
5:     sum = 0
6:     for j = 0; j < n; j++ do
7:         sum = sum + Train[j][i]
8:     end for
9:     promedio[i] = sum/n
10: end for
11: \\ calculamos la matriz de covarianza:
12: for i = 0; i < n; i++ do
13:     for j = 0; j < 784; j++ do
14:         Train[i][j] = Train[i][j] - promedio[j]
15:     end for
16:     Train[i][j] = Train[i][j] /  $\sqrt{n-1}$ 
17: end for
18: matriz covarianza  $\leftarrow$  Traint * Train
19: \\ obtenemos la matriz con los  $\alpha$  autovectores de la matriz de covarianza como columna (P):
20: P  $\leftarrow$  metodo de la potencia y de deflacion (covarianza,  $\alpha$ )
21: for i = 0; i < cantidad de imagenes de test; i++ do
22:     for j = 0; j < 784; j++ do
23:         test[i][j] = test[i][j] - promedio[j]
24:     end for
25:     test[i][j] = test[i][j] /  $\sqrt{n-1}$ 
26: end for
27: \\ realizamos la transformación característica para train (TcTrain) y para test (TcTest):
28: TcTrain  $\leftarrow$  Train * P
29: TcTest  $\leftarrow$  Test * P
30: \\ finalmente procedemos a reconocer los dígitos:
31: Knn(TcTrain, TcTest)

```

2.3. K-Fold Cross Validation

Para poder concentrarnos en la evaluación de los métodos y la óptima elección de los parámetros, necesitamos probar los mismos sobre la base de train, ya que sobre sus elementos conocemos el dígito al que pertenece cada uno. Ante esta situación, particionaremos dicha base en dos, utilizando una parte de ella en forma completa para el training y la restante como test, pudiendo así verificar la predicción realizada y mostrar las tasas de efectividad (La cantidad de dígitos correctamente clasificados respecto a la cantidad total de casos analizados).

Sin embargo, surge un problema no menor, realizar la experimentación sobre una única partición podría traer como consecuencia predicciones no deseadas, y una mala estimación de los parámetros, por ejemplo podría ocasionar *overfittin*, que el algoritmo del método bajo análisis quede ajustado a una serie de características muy específicas de la base de train que no tienen relación con el objeto de estudio. Para solucionar el problema en cuestión, se usará la técnica de Cross Validation, en particular el K-Fold Cross Validation, para que los resultados de la experimentación resulte estadísticamente más robustos.

La técnica de K-Fold Cross Validation consiste en particionar de forma aleatoria la base de train en K conjuntos de igual tamaño (Siempre asumiendo que el cardinal del conjunto de la base de train es múltiplo de K). Luego se realizan K ejecuciones del programa, cada una de ellas eligiendo un conjunto para test, y utilizando los K-1 conjuntos restantes para train. Para las K ejecuciones se usarán los mismos parámetros en cada una de ellas. Además, en el método tradicional se suelen realizar varias corridas para un mismo valor de K.

Para calcular estas particiones usaremos el comando CVPARTITION de MATLAB de la siguiente manera:

Algoritmo 3 calcularParticion(int K , int n)

```
1: C = cvpartition(n,'Kfold',K)
2: fileID = fopen('K.txt','w');
3: for int i = 1:K do
4:     fprintf(fileID,'d ',C.training(i))
5:     fprintf('n')
6: end for
7: fclose(fileID)
```

2.4. Evaluación de los métodos

Para comprobar la eficacia del método KNN y KNN + PCA procederemos a desarrollar distintas experimentaciones. En primer lugar para KNN variaremos la cantidad de vecinos (k) que vamos a utilizar para reconocer a que clase pertenece cada imagen de test. Luego, para PCA + KNN también vamos a variar la cantidad de vecinos a tomar, fijando un α (cantidad de componentes principales), y para el mejor de ellos (evaluando calidad de la solución y tiempo de cómputo) procederemos a variar el valor de α utilizando el mejor k obtenido anteriormente. Los valores obtenidos por estas experimentaciones serán detallados en la sección resultados, los mismos fueron obtenidos en una pc Intel Core i3 con 4gb de RAM.

2.5. Elección de K, hipótesis 1

Para las siguientes experimentaciones aplicaremos la técnica de *K - Fold Cross Validation*, para la misma es necesario la elección de un valor K que representa la cantidad de conjuntos en la que dividiremos la base de train. Dado que para la formación de las particiones, que serán utilizadas tanto para test como para train, utilizamos el comando CVPARTITION la misma las calcula como:

cantidad de imágenes de test = n/K (1)

cantidad de imágenes de train = $n - n/K$ (2)

con n la cantidad de imágenes total, en nuestro caso las 42.000 de la base de train.

A partir de esta distribución elaboramos una primer hipótesis.

- Hipótesis N° 1: Para los K de mayor valor, la tasa de reconocimiento promedio es mayor que para los de menor valor.

Como se menciono, lo que nos llevo a formular esta hipótesis es la distribución en la cantidad de imágenes de test y de train al utilizar distintos K . Notemos por (1) que la cantidad de imágenes utilizadas como test aumenta a medida que K disminuye y como consecuencia las de train disminuyen (2).

Si al momento de reconocer un dígito i , se tienen pocas imágenes con las que comparar (train) entonces tenemos menos representaciones de ese dígito, por lo que al momento de elegir los k vecinos estamos mas propensos a que solo una fracción de los mismos pertenezcan a la misma clase que i . Como en la elección de la clase, a la que supuestamente pertenecería i , también intervienen los otros vecinos (los que tienen clase distinta a la de i) podría ser que la clase prioritaria entre los k vecinos no sea la de i , dado a que hay pocas representaciones de este dígito, terminando en un reconocimiento erróneo. Es por esto que creemos que, en contraste, si aumentamos las imágenes de train habria mayores probabilidades de que la mayoría de los vecinos sean de la misma clase que i , obteniendo una mayor tasa de reconocimiento.

Para poder corroborar esta hipótesis a las experimentaciones mencionadas anteriormente (variación de k , y α) también las haremos variando el valor de K . Se eligieron tres valores para experimentar, los mismos son $K = 2, 10$, y 20 , dicha elección esta ligada a como quedan distribuidas la cantidad de imágenes utilizadas tanto para test como para train. Cuando $K = 2$ solo un 50 % del total de las imágenes son de entrenamiento (además es el mínimo valor permitido por matlab para el comando cvpartition, usando Kfold); para $K = 10$ un 90 % y para 20 un 95 %, de esta manera abarcamos un amplio y variado rango de valores para la cantidad de imágenes a reconocer y las utilizadas como entrenamiento. Como consecuencia de esta elección surge otro problema a analizar y es que, si al aumentar el K las tasas de reconocimiento promedio aumentan ¿que porcentaje de mejora obtenemos (si es que lo hacen)? y ¿como influye el tiempo de cómputo en estas variaciones?. Por lo que en las experimentaciones además buscaremos responder estas preguntas para poder determinar la mejor relación entre calidad de soluciones y tiempo.

2.6. Experimento 1: Variando el k en kNN

Una de las experimentaciones a desarrollar es analizar el comportamiento de la técnica KNN para distintos valores de k . Una hipótesis que podemos elaborar en esta experimentación es:

- Hipótesis N° 2: En el método de kNN, al incrementar k vamos a obtener una menor tasa de efectividad. En otras palabras, se resumiría que a mayor valor de k , menor porcentaje de la tasa.

La elaboración de esta hipótesis se basa en que al aumentar el valor de k , para clasificar la clase de un dígito i , se deben tener en cuenta más vecinos pudiendo aumentar la cantidad de dígitos distintos de i a tal punto que la clase de otro dígito sea la que predomine. Por ejemplo si entre los k vecinos hay X que pertenecen a la clase de i e Y que pertenecen a una clase j (distinta de i) con: $X < Y$, $X + Y \leq k$.

Entonces se elegiría la clase j antes que la de i porque es la que predomina dentro de los k vecinos seleccionados pero, sería errónea la elección. En cambio si se eligiera un k menor y adecuado, la cantidad de imágenes de la clase j podría ser menor a la de i , y de esta manera se seleccionaría la clase correcta. Para poder corroborar nuestra hipótesis tomaremos los valores $K = 2, 30, 100$ y 500 . De esta manera para los tres primeros valores deberían haber diferencias entre las tasas de reconocimientos pero, no tan abruptas que cuando le damos a k un valor mucho mas grande, en comparación a los elegidos anteriormente que es cuando $k = 500$. Además aplicaremos la técnica de K-Fold cross validation, con los valores de K definidos en la sección *Elección de K , hipótesis*.

2.7. Experimento 2: PCA + KNN

2.7.1. Variando k

Dados los valores de K elegidos previamente, realizaremos una serie de mediciones de tiempos y tasas para $k \in \{2, 30, 100, 500\}$, fijando un valor de $\alpha = 50$. El objetivo de estas mediciones es conseguir el valor de k que mejor se comporte de manera que estén equilibrados el tiempo de cómputo y la tasa de reconocimiento. Luego, tomaremos el valor de k conseguido y procederemos a variar α . Los valores de k elegidos representan un subconjunto de los que se utilizaron en el experimento sobre KNN con el objetivo de que luego podamos comparar las tasas de reconocimiento y el tiempo de cómputo que producen ambos métodos. El valor de α fue fijado en 50 porque, si bien es un valor relativamente chico para el rango en el que se encuentra (de 1 a 784), observamos que las tasas de reconocimiento con este valor son sorprendentemente altas. Esto significa que no tendría sentido aumentar α para obtener (o no) una mínima diferencia en la tasa de reconocimiento y terminar pagando un costo de cómputo y memoria (a mayor α mayor son las dimensiones de las matrices que utilizamos) mucho mayor.

2.7.2. Variando α

Una vez ya conseguido el valor de k que buscamos, la idea ahora es probar distintos valores de α y ver si al tomar valores grandes obtenemos tasas de reconocimiento mayores (esta será nuestra tercera hipótesis del trabajo), fomalizando:

- Hipótesis N° 3: En el método de kNN+PCA, al incrementar α vamos a obtener una mayor tasa de efectividad. En otras palabras, se resumiría que a mayor valor de α , mayor porcentaje de la tasa.

En principio parece muy razonable creer que esto va a suceder ya que al aumentar α estamos agrandando la nueva base en la cuál vamos a expresar nuestros datos. O sea cuanto más grande es α , menor es la cantidad de información que "recortamos" de cada muestra. Los valores de α que vamos a elegir van a ser los del siguiente conjunto: $\{20, 100, 300\}$, no contamos el $\alpha = 50$ porque ya lo usamos cuando variamos k pero también tenemos ese resultado. El máximo que usamos es

$\alpha = 300$ debido a que, como estamos utilizando PCA con el objetivo de reducir el tiempo de cómputo de KNN, en la práctica vamos a tratar de no utilizar valores muy grandes y 300 es un valor que está cerca de la mitad en el rango de valores que puede llegar a tomar α . El valor 20 lo elegimos para ver si al reducir α la tasa de reconocimiento empeora así como el 100 es elegido para ver si aumentar al doble el valor de α nos da una tasa de reconocimiento mucho mayor y, si esto no sucede entonces vemos que pasa con un valor mucho más grande como puede ser 300.

3. Resultados

3.1. Experimento 1

Se procedió a ejecutar el método de kNN variando k, para $K = 2, 10$ y 20 con $k \in \{2, 30, 100, 500\}$.

Se observó que la memoria utilizada por el programa era de 127,6 Mb en sus fases iniciales, hasta que en un punto se queda en 380,5 Mb.

Los resultados obtenidos fueron los siguientes:

- Con $K = 2$

k	Tiempo	Tasa
2	15941.152	0.952476
30	19450.181	0.942452
100	19116.135	0.91681
500	20496.988	0.844905

Tabla 1: Tasas para Knn variando k para dos particiones

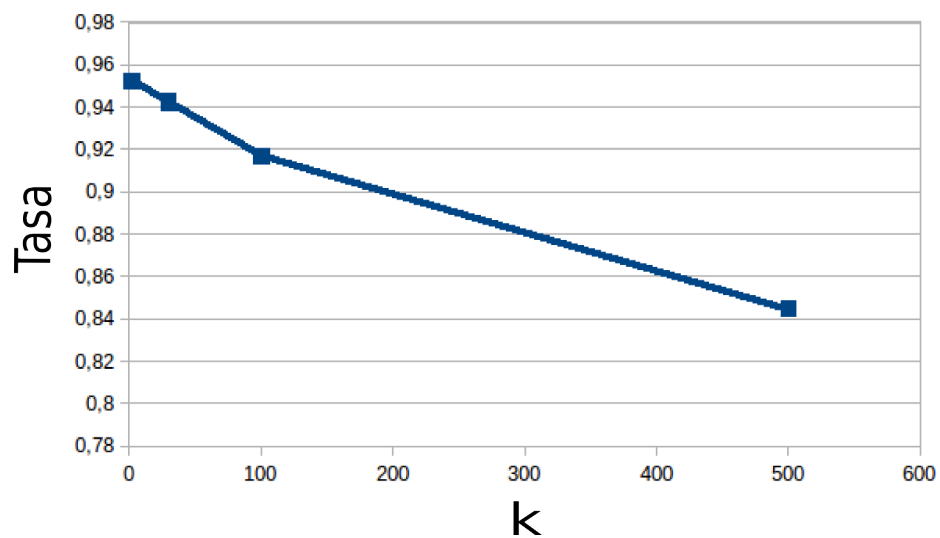


Figura 1: Tasas para Knn variando k para dos particiones

- Con $K=10$

k	Tiempo	Tasa
2	31056.21	0.96169
30	34590.4	0.952262
100	34449.695	0.93169
500	37274.029	0.878833

Tabla 2: Resultados para Knn variando k para diez particiones

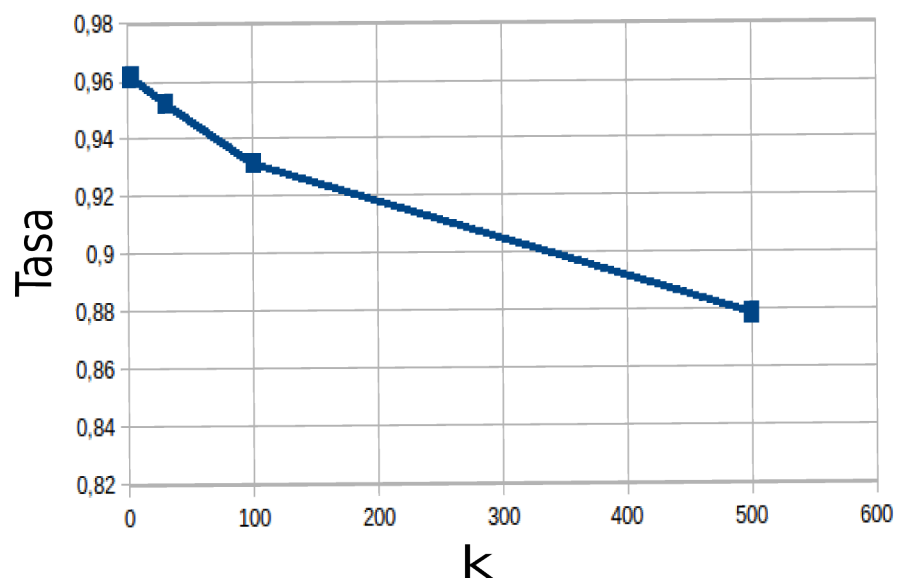


Figura 2: Tasas para Knn variando k para diez particiones

- Con $K = 20$

k	Tiempo	Tasa
2	238893.923	0.962381
30	263245.053	0.953167
100	284027.496	0.933429
500	313345.443	0.881381

Tabla 3: Tasas para Knn variando k para veinte particiones

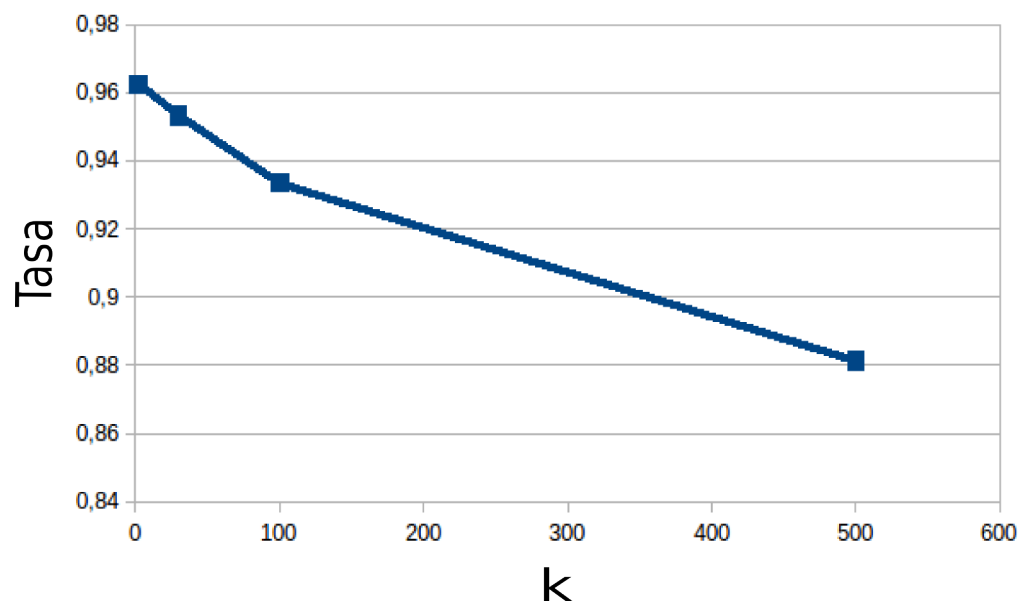


Figura 3: Tasas para Knn variando k para veinte particiones

En base a estos resultados se puede apreciar claramente la caída de la tasa en función de la cantidad de vecinos mas cercanos(k), y el aumento de tiempo con el mismo.

3.2. Experimento 2

3.2.1. Variando k

A continuación, presentamos las tablas correspondientes a las tasas de reconocimiento y el tiempo de cómputo para los valores de k elegidos en el desarrollo y para los distintos valores de K :

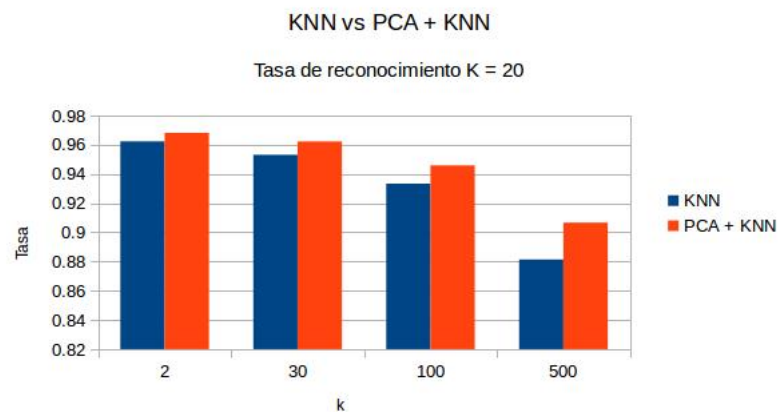
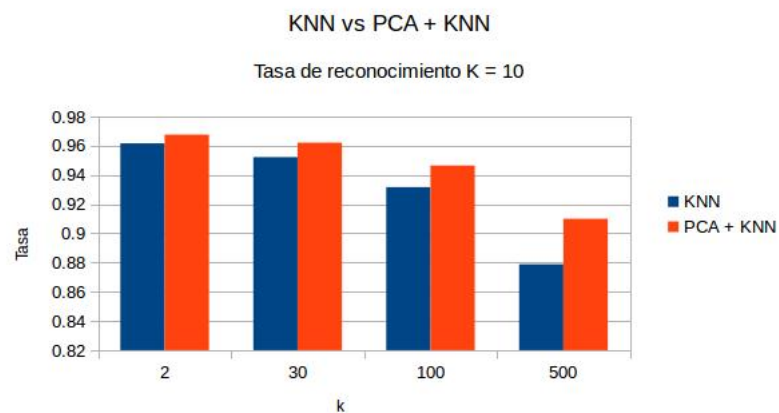
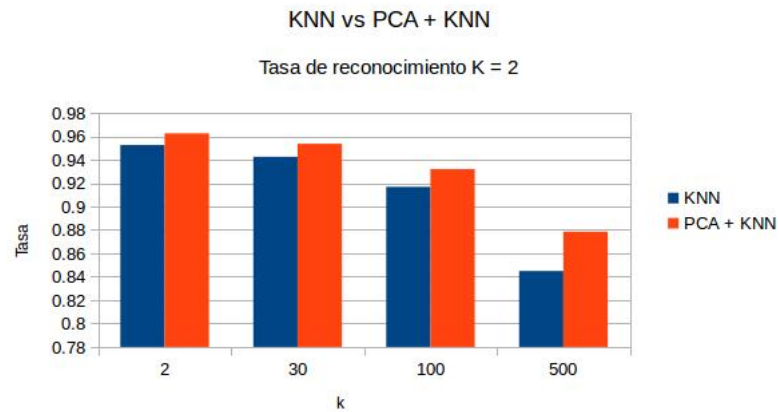
Tasas:

k	K = 2	K = 10	K = 20
2	0.9625	0.96769	0.968214
30	0.953595	0.962143	0.962357
100	0.931976	0.946493	0.945905
500	0.878476	0.910008	0.906643

Tiempos:

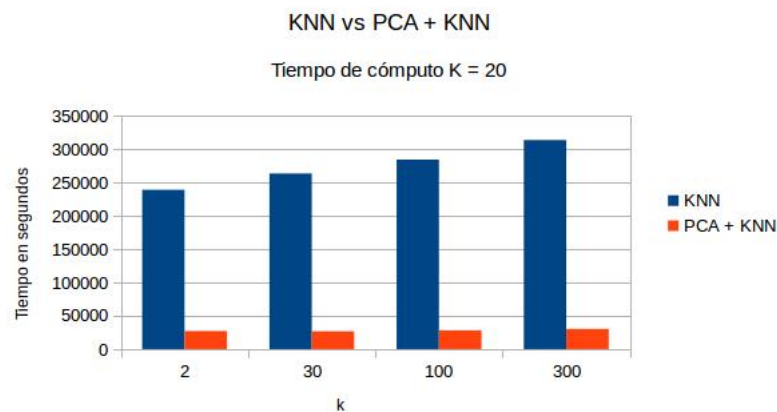
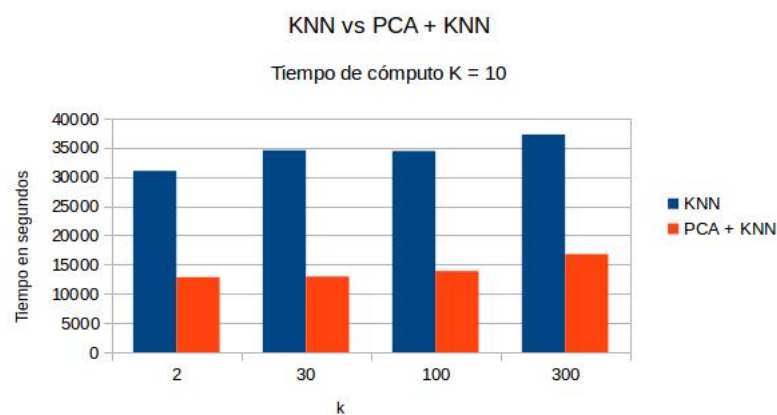
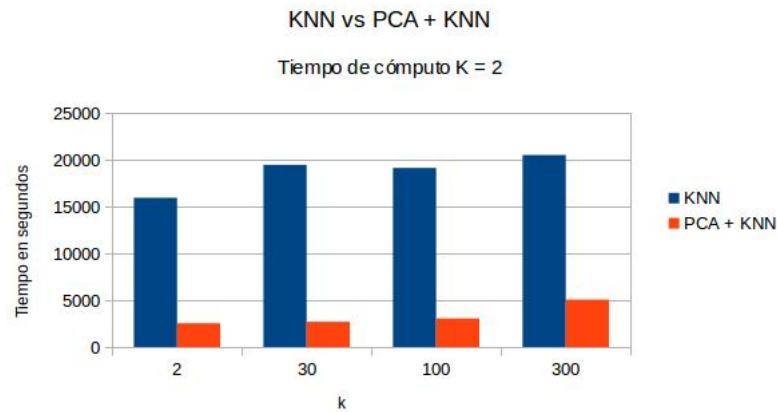
k	K = 2	K = 10	K = 20
2	6717.76	12842.6	27005.4
30	7167.14	12968.6	26608
100	8099.62	13892.1	28122.2
500	13445.4	16788.3	30181.4

Ahora vamos a comparar las tasas de PCA + KNN con las de KNN para los distintos K :



Podemos observar como sorprendentemente PCA siempre supera a KNN, teóricamente esto no debería suceder. Aunque hipotéticamente se deba a un arrastre de error por estar trabajando con aritmética finita, o por las característica de las imágenes, que al aplicar PCA, este me elimine los datos que distorcionan de manera muy grande la muestra. Es decir que PCA, aun siendo chico el valor α me capture gran parte de los datos esenciales de cada imagen.

A continuación presentamos un gráfico que compara los tiempos de KNN y PCA + KNN para $K = 2$ y $K = 10$, fijando $\alpha = 50$ en PCA y variando k :



En los gráficos muestran lo esperable que es que los tiempos de KNN son más grandes que los de PCA + KNN. Para los gráficos sobre $K = 20$, el tiempo de cómputo fue calculado de manera aproximada (multiplicando por 20 el tiempo que se tarda en procesar una partición de las 20) debido a que es demasiado grande. La tasa de reconocimiento en este caso tampoco es el promedio de las tasas porque no las tenemos, si no que es la tasa de reconocimiento de la partición particular que procesamos.

3.2.2. Variando α

A continuación, presentaremos la tablas de los distintos K para los cuales corrimos PCA + KNN variando α . En estos experimentos siempre vamos a estar utilizando el mismo valor de k que va a ser 2 ya que fue elegido por ser el que maximiza la tasa de reconocimiento y minimiza el

tiempo de cómputo.

Alfa	20	50	100	300
K = 2	0.95269	0.962524	0.95831	0.956952
K = 10	0.957619	0.96769	0.965524	0.965214
K = 20	0.958595	0.968214	0.966524	0.965929

Vemos como no se cumple la hipótesis que planteamos de que si aumentamos el valor de α entonces va a aumentar la tasa de reconocimiento ya que en los tres casos la tasa más alta está asociada a $\alpha = 50$ y es estrictamente mayor a las tasas con $\alpha = 100$ y $\alpha = 300$.

Veamos los tiempos de cómputo asociados a estos valores de α :

	Alfa = 20	Alfa = 50	Alfa = 100	Alfa = 300
K = 2	2524	2693	3043	5052
K = 10	11234	12842.6	16167.8	34122.9
K = 20	27604.4	27005.4	33377	61360.8

Se puede ver en el gráfico como al hacer un aumento significativo en el α el tiempo de cómputo puede llegar a aumentar bastante. En los tres casos vemos que trabajar con $\alpha = 300$ requiere un tiempo de cómputo cercano al doble de lo que requiere con $\alpha = 20$. Esto significa que para valores cercanos de α el tiempo de cómputo no se va a ver afectado de manera significativa pero si realizamos aumentos grandes entonces es probable que el tiempo aumente de manera lineal.

Dado que nosotros queremos encontrar la mejor configuración tanto a nivel calidad de los resultados como tiempo de cómputo y vimos que las tasas variando α en 20, 50 100 y 300 son muy parecidas podríamos elegir $\alpha = 20$. De esta manera estaríamos minimizando el tiempo de cómputo y obteniendo una tasa de reconocimiento bastante alta. Sin embargo, en un contexto donde el tiempo de cómputo puede llegar a ser mucho más grande, por ejemplo porque estamos trabajando con un K mucho mayor, porque contamos con bases de datos de mayor tamaño o por cualquier otro motivo, sería interesante poder reducir lo máximo que podamos al valor de α . Es decir, nosotros sabemos que las tasas con $\alpha = 20, 50, 100$ y 300 son muy similares, pero ¿cuanto más podemos disminuir α sin que la tasa de reconocimiento baje demasiado?. Para contestar esta pregunta decidimos disminuir el valor de α de una manera más drástica y ver que sucede. A continuación, se presentan las tasas de reconocimiento asociadas a los distintos K , para $\alpha = 5$:

K	Tasa de rec.
2	0.667595
10	0.651357
20	0.650857

Los resultados demuestran que la diferencia de las tasas con respecto a los valores anteriores de α es claramente notable, lo que significa que deberíamos elegir un α entre 5 y 20. Es importante observar como al realizar una variación en el valor de α , las tasas para los distintos valores de K se mantienen muy parecidas. Esto significa que para reducir el tiempo de cómputo de la experimentación, podemos a tratar de encontrar este α utilizando $K = 2$ y lo que obtengamos también nos va a servir para $K = 10$ y $K = 20$. Probemos con $\alpha = 7, 13$ y 16.

Alfa	Tasa de rec.
7	0.821143
13	0.926857
16	0.944714

Bien, la tabla muestra que para $\alpha = 13$ la tasa es ≈ 0.92 . Dado que con un valor de 7 estamos pasando el 0.8 y con 13 estamos pasando 0.92, tomando un valor de $\alpha = 10$ nuestra tasa va a ser ≈ 0.9 , lo cual es más que aceptable. Partimos de un valor de $\alpha = 50$ y experimentando, llegamos a que si bajamos ese valor a 10 entonces vamos a tener una tasa un poco peor pero igual muy buena y vamos a estar reduciendo mucho más la dimensión de nuestras muestras.

4. Discusión

4.1. Hipótesis 1:

Con respecto a la hipótesis 1, que establecía que corriendo cualquiera de los 2 métodos implementados (kNN y kNN+PCA) fijando k y α , a mayor valor de K mayor porcentaje de tasas, procederemos primero a observar los resultados de kNN+PCA.

Comparando las tasas de efectividad resultantes de usar $K=2$ contra $K=10$ y $K=20$, se nota claramente que las de $K=2$ son menores que las de los otros dos valores de K . Esto es consecuencia principal del uso de la técnica de K-Fold Cross Validation, ya que con un valor K chico nuestra partición de la base de entrenamiento se divide en partes mas grandes, en consecuencia la longitud de la base de test tiende a ser igual a la base de entrenamiento. Dicho en otros términos, tenemos menos elementos en la base de entrenamiento para entrenar nuestro algoritmo y menos permutaciones sobre cuales elementos los consideramos parte del test. Esto es la principal causa de que con $K=2$ obtengamos los menores valores de la tasa de efectividad. Para que la tasa de efectividad tienda a 1, es necesario contar con una base lo más extensa posible. Y además al aumentar K , tenemos mas combinacion de particiones posibles y un entrenamiento más profundo. En este caso, la disminucion de la misma no se ve afectada por el valor que tome k y α .

Hasta aqui la hipótesis se confirmaría, pero al analizar y comparar las tasas de $K=10$ contra $K=20$, observamos que aca si incide el valor que tome k . Para k con valores chicos, se ve una diferencia de la tasa cuando saltamos de $K=10$ a $K=20$, pero a medida que k aumenta, las tasas empiezan a comportarse de manera similar, teniendo vaivenes para algunos k , en el que la tasa de $K=10$ le gana a $K=20$ o viceversa. Ante esta situación, decidimos observar más detalladamente los datos de la experimentación y logramos ver que para k grandes y $K=20$, la muestra de las tasas de cada partición posee una mayor varianza, es decir que los valores tienen mayores fluctuaciones en comparación a la media de la muestra. Realizando un análisis más intensivo, deducimos que esto se debe a que a mayor valor de K , tengo mas elementos de entrenamiento, y a su vez a mayor valor de k , tengo más elemento que entrarán en la definición de vecinos más cercanos, pudiendo aqui encontrar elementos que no son del mismo label que el elemento en estudio y así alterar notablemente la tasa de efectividad en cada una de las particiones.

Como último dato, el valor de α no incide en las tasas de efectividad entre diferentes K , es decir, para un α fijo, a medida que se aumenta el valor de K , aumenta la tasa.

Con respecto al método de kNN, la elección de los K , se comporta de la misma manera y tiene las mismas consecuencias que en el método de kNN+PCA. Finalmente se podría concluir que la hipótesis queda parcialmente verificada, ya que la única parte que se refuto de la misma fue la parte de mantener k constante. Sobre esta particularidad de k , se continuará detallando en la discusión del experimento 1.

Análisis de tiempo de computo al variar K : En el método de kNN+PCA, claramente se incrementa el tiempo de cómputo a medida que aumenta el valor de K , esto se debe a que el cálculo de la matriz de covarianza del entrenamiento domina gran parte del tiempo del metodo (sin contar la llamada a kNN). Dicha matriz se calcula de la siguiente forma: por ejemplo, si $K=2$, la matriz de entranamiento va a tener un tamaño de 21000×784 , su transpuesta de 784×21000 , entonces la matriz de covarianza se calcula multiplicando la transpuesta por la matriz de entrenamiento resultando siempre una matriz de 784×784 . Para este caso de K , haremos 2 multiplicaciones de matrices de 784×21000 y 21000×784 . Ahora bien, si suponemos que $K=10$, el procedimiento es el mismo, salvo que ahora, mi base de entrenamiento va a ser mas grande, por ende, su matriz también que en este caso va a tener tamaño de $(42.000-42.000/10) \times 784 = 37800 \times 784$. Para este caso de K , haremos 10 multiplicaciones de matrices de 784×37.800 y 37.800×784 . Al aumentar el valor de K , además de hacer más multiplicaciones por tener más particiones, las matrices aumentan de tamaño por lo explicado anteriormente. Para $K=20$ es el mismo procedimiento.

En el método de kNN, por ejemplo para $K=2$, voy a recorrer 2 veces la base de entrenamiento de 21000 dígitos para un base de test con la misma cantidad de dígitos. Para $k=10$, voy a recorrer 10 veces la base de train de 37.800 dígitos para una base de test de 4.200 digitos. Realizando un análisis más exhaustivo, para el primer caso de $K=2$, estaríamos recorriendo 21.000 veces la base de 21.000 digitos, todo eso 2 veces lo que resultaria en $2 \times 21.000 \times 21.000 = 822.000.000$ iteraciones.

Para el caso de $K=10$, estaríamos recorriendo 4200 veces la base de 37800 dígitos, todo eso 10 veces lo que resultaría en $10 \cdot 4.200 \cdot 37.800 = 1.587.600.000$ iteraciones, lo que implicaría un 93 % más de iteraciones. Para $K=20$ pasa lo mismo.

4.2. Experimento 1:

Podemos ver que en ningún caso se obtiene una tasa del 100 %. Con un k chico se corre el riesgo de que los vecinos más cercanos no sean los correctos por una situación especial que pueda llegar a ocurrir, y entonces el dígito no sea reconocido, por ejemplo que justo el dígito en estudio este rodeado de k dígitos con otra caracterización. Se verifica empíricamente la hipótesis N° 2, con un k grande las tasas disminuyen apreciablemente. Esto se lo podemos atribuir a que hay una probabilidad no despreciable de que un dígito a reconocer este mas cercano a un dígito con distinta caracterización que a uno con igual. No es una cuestión de falta de dígitos iguales al que se esta reconociendo. Por ejemplo al tomar 500 vecinos más cercanos, no es que no halla sido posible tomar a estos 500 como el dígito a reconocer por falta de este ultimo en la matriz de train, sino que en lugar de tomarse un dígito con igual caracterización, se toma uno con distinta, pero más cercano.

También podemos ver que al incrementar k se obtiene un tiempo mayor. Esto se debe a que el algoritmo de knn en cada paso, recorre un vector de tamaño k donde se encuentran los mas cercanos, y lo va ir actualizando.

4.3. Experimento 2:

4.3.1. Variando k

Al realizar las variaciones de k propuestas en el desarrollo obtuvimos distintas tasas de reconocimiento que, como podemos observar en la primera tabla, van empeorando a medida que aumenta el k . Esto pasa porque, aunque estemos haciendo PCA primero, luego vamos a estar aplicando KNN y ya vimos que cuando aplicamos KNN y aumentamos el valor de K , las tasas empeoran, independientemente del preprocesamiento de los datos.

4.3.2. Variando α

En los experimentos vimos que nuestra hipótesis de que al aumentar el valor de α íbamos a estar aumentando la tasa de reconocimiento es falsa. Sin embargo, cuando bajamos de $\alpha = 50$ a $\alpha = 5$ pudimos notar una diferencia importante en la tasa de reconocimiento. Una posible explicación a lo que está sucediendo es que, si bien matemáticamente tiene sentido pensar que deberíamos tener tasas de reconocimiento más altas al aumentar α (porque agrandamos nuestra base en la cual expresamos nuestros datos), nosotros estamos implementando todo esto en una computadora. Esto significa que todo lo que hacemos tiene cierto error y que ese error puede ir incrementandose.

Los autovectores por ejemplo, los calculamos mediante el método de la potencia, pero no debemos olvidarnos que el método de la potencia se basa en aproximar un límite. Esto quiere decir que los autovalores y autovectores que estamos calculando, en realidad no van a ser exactamente los que queremos obtener. En la experimentación vimos que a partir de un cierto valor, ya no tiene sentido seguir aumentando α porque las tasas dan muy similares (incluso peores) y el tiempo de cómputo aumenta. Tranquilamente lo que podría estar pasando es que, los primeros autovectores que estemos calculando tengan menos error (porque el método de la potencia para ellos converge más rápidamente) y a medida que se va aumentando la cantidad de autovectores, el error que se produce al calcular cada uno de ellos es cada vez más grande. Esto explicaría lo que pasa en nuestros resultados porque los primeros autovectores de la base son los que estarían aportando más información (porque son los que menos error tendrían) y a partir de un punto, el error es tan grande que ya no aporta información confiable y puede ocurrir entonces que la tasa de reconocimiento empeore.

4.4. Competencia Kaggle

Para realizar nuestro submission en la página de Kaggle, a la competencia de Digit Recognizer, modificamos levemente el código de nuestro programa. El mismo se encuentra en la carpeta Kaggle. Las modificaciones se deben a que en este caso, la base de test proporcionada por la página no tiene labels, no utilizamos el método de K-Fold Cross Validation, la función kNN ahora tiene que devolver las predicciones de los dígitos, entre otras cosas.

Luego de toda la experimentación concluimos que los mejores parámetros, en cuanto tiempo y eficacia, para correr el programa son algún k y α cercanos a:

- $\alpha = 50$
- $k = 2$

A continuación seguiremos experimentando y probaremos con valores cercanos a estos parámetros para estudiar la efectividad que nos proporciona el sistema de Kaggle.

- Con $k=2$ y $\alpha=50$, tuvimos una efectividad de 0.96686 con un tiempo de computo de 1338.35
- Con $k=3$ y $\alpha=50$, tuvimos una efectividad de 0.96857 con un tiempo de computo de 1454.45.
- Con $k=3$ usando solo kNN, tuvimos una efectividad de 0.97200 con un tiempo de computo de 12691.6.
- Con $k=4$ y $\alpha=50$, tuvimos una efectividad de 0.97286 con un tiempo de computo de 1654.71.
- Con $k=5$ y $\alpha=50$, tuvimos una efectividad de 0.97286 con un tiempo de computo de 1776.28.
- Con $k=4$ y $\alpha=60$, tuvimos una efectividad de 0.97314 con un tiempo de computo de 1950.58.
- Con $k=10$ y $\alpha=70$, tuvimos una efectividad de 0.97086 con un tiempo de computo de 3050.58.
- Con $k=1$ y $\alpha=60$, tuvimos una efectividad de 0.97200 con un tiempo de computo de 1750.48

5. Conclusiones

En el presente trabajo pudimos apreciar que la utilización de los métodos KNN y PCA, para ciertas configuraciones (de α y k), nos proveen tasas de reconocimiento bastante altas. Dichas configuraciones requieren tiempos de cómputo muy razonables ya que vimos que aumentar demasiado α no tiene sentido y que a valores muy grandes de k la tasa disminuye. Es decir, obtener tasas de reconocimiento altas implica utilizar parámetros de k y α relativamente chicos y como el tiempo de cómputo se ve afectado negativamente cuando estos parámetros son grandes, entonces los tiempos asociados a tasas altas no son necesariamente grandes.

Sin embargo, el método de KNN utilizando la norma dos como distancia entre dígitos tiene un margen de error para reconocer a los mismos, lo cual hace que en algunas situaciones no se pueda utilizar, ya que no es aceptable una equivocación. El error de este último método se debe en parte a la cercanía (en norma dos) de dígitos distintos, pero también en parte a los errores numéricos que se propagan al hacer operaciones aritméticas en la computadora.

PCA es un método que resulta muy efectivo para el reconocimiento de dígitos por varios motivos. En teoría deberían obtenerse tasas menores que con KNN, ya que se descarta información cuando se reduce la dimensión. Sin embargo, PCA arrastra menos error numérico y en consecuencia en nuestros experimentos obtenemos tasas mayores. Hipotéticamente, también pensamos que esta situación se deba a que PCA, aun así siendo el α no tan grande, los datos que capture sean la mayoría de los datos esenciales de cada imagen, descartando los datos que traerían ruido al procesamiento con kNN. En un futuro, nos gustaría experimentar más a fondo, para determinar cual de estas dos situaciones es realmente la que pasa, o cual es el grado de cupabilidad de cada una.

Por otro lado, este método al tratar con menos cantidades de coordenadas para las imágenes el tiempo de cómputo es mucho menor.

Utilizando cross validation pudimos realizar experimentos cuyos resultados nos proveen información objetiva, ya que las particiones son generadas de manera aleatoria y la tasa final de un método para cierta configuración es el promedio de las tasas de cada partición. En algunos casos, al trabajar con valores de K muy grandes tuvimos que realizar aproximaciones sobre el tiempo de cómputo total que tardaría la computadora en procesar todas estas particiones y utilizar la tasa de reconocimiento de una partición sola. De todas formas, esta información si bien está aproximada, nos permite ver como se puede agrandar la brecha entre el método de KNN y PCA + KNN para valores de K mayores.

Sería interesante realizar, en un futuro, experimentos que permitan ver para que configuraciones KNN es mejor que PCA + KNN en la práctica ya que en nuestros resultados siempre es mejor PCA + KNN, tanto en tiempo de cómputo como en tasas de reconocimiento. Para esto, deberíamos trabajar con una cantidad de valores de α y k más grande. También queda pendiente considerar otras normas para las distancias entre las imágenes, y ver que tan efectivo es el método. Además, en el marco de integración de todas las materias de la carrera, nos gustaria probar si se puede acelerar el computo de operacion entre vectores o matrices, usando lo que aprendimos en Organización del Computador II sobre procesamiento paralelo usando SIMD en procesadores con arquitectura Intel x86.

6. Apéndices

Apéndice A:

Métodos Numéricos
Primer Cuatrimestre 2015
Trabajo Práctico 2



Departamento de Computación
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Introducción

Uno de los momentos de mayor tensión durante un cuatrimestre se genera en épocas de parciales. El día del examen, luego de entregar la resolución del mismo, suele aparecer esa inmanejable incertidumbre en los alumnos: *¿entregué todas las hojas?*, *¿consideré el caso particular de $\beta = 0$?*, *¿por qué mi compañero puso verdadero, si yo encontré un contraejemplo?* Desde el lado del docente, la dificultad radica en corregir un gran volumen de ejercicios técnicamente complejos en un tiempo acotado, buscando deducir en base a información limitada si el alumno comprendió o no el tema.

Estos dos factores llevaron a la conformación de una comisión interclaustró de docentes y alumnos, para quienes resguardamos su identidad, a proponer como solución el desarrollo de una herramienta automática de corrección de exámenes, con el fin de reducir los tiempos en la devolución de los mismos. Para evitar suspicacias respecto de las correcciones, se propuso dividir el problema en varias etapas y que el sistema sea desarrollado por alumnos, contando desde ya con la validación de los docentes. La primera etapa del proyecto consiste en desarrollar una herramienta que permita automáticamente identificar dígitos manuscritos, a ser aplicada a una digitalización de cada examen, y cuyo post-procesamiento involucrará darle semántica y reglas a la información obtenida.

Como es de esperar, el éxito de la herramienta depende inevitablemente de disponer de un buen reconocedor de dígitos manuscritos. Para semejante responsabilidad, dado el gran nivel mostrado durante el comienzo del cuatrimestre, la comisión asesora ha decidido elegir a los alumnos de Métodos Numéricos para llevar adelante el desarrollo.

Objetivos y Metodología

Como instancias de entrenamiento, se tiene una base de datos de n imágenes de $M \times M$ píxeles, las cuales se encuentran etiquetadas con el dígito, 0-9, al que corresponden. Definimos como $m = M \times M$ al número total de píxeles de una imagen. Asumiremos también que las imágenes se encuentran en escala de grises (cada pixel corresponde a un valor entre 0 y 255, indicando la intensidad del mismo) y que el etiquetado no contiene errores. El objetivo del trabajo consiste en utilizar la información de la base de datos para, dada una nueva imagen de un dígito sin etiquetar, determinar a cuál corresponde teniendo en cuenta factores de calidad y tiempo de ejecución requeridos.

Una primera aproximación es utilizar el conocido algoritmo de *k Vecinos más Cercanos* (*kNN*, por su nombre en inglés). En su versión más simple, este algoritmo considera a cada objeto de la base de entrenamiento como un punto en el espacio, para el cual se conoce a qué clase corresponde (en nuestro caso, qué dígito es). Luego, al obtener un nuevo objeto que se busca clasificar simplemente se buscan los k vecinos más cercanos y se le asigna la clase que posea el mayor número de repeticiones dentro de ese subconjunto, es decir, la moda. Con este objetivo, podemos representar a cada imagen de nuestra base de datos como un vector $x_i \in \mathbb{R}^m$, con $i = 1, \dots, n$, y de forma análoga interpretar las imágenes a clasificar mediante el algoritmo *kNN*.

Sin embargo, el algoritmo del vecino más cercano es sensible a la dimensión de los objetos a considerar y, aún aplicando técnicas de preprocesamiento, puede resultar muy costoso de computar. Teniendo en cuenta esto, una alternativa interesante de preprocesamiento que busca reducir el tamaño de los elementos a considerar es el método de *Análisis de Componentes Principales* (PCA, por su sigla en inglés), que consiste en lo siguiente. Para $i = 1, \dots, n$, recordar que $x_i \in \mathbb{R}^m$ la i -

ésima imagen de nuestra base de datos almacenada por filas en un vector, y sea $\mu = (x_1 + \dots + x_n)/n$ el promedio de las imágenes. Definimos $X \in \mathbb{R}^{n \times m}$ como la matriz que contiene en la i -ésima fila al vector $(x_i - \mu)^t / \sqrt{n-1}$, y

$$A = U \Sigma V^t$$

a su descomposición en valores singulares, con $U \in \mathbb{R}^{n \times n}$ y $V \in \mathbb{R}^{m \times m}$ matrices ortogonales, y $\Sigma \in \mathbb{R}^{n \times m}$ la matriz diagonal conteniendo en la posición (i, i) al i -ésimo valor singular σ_i . Siendo v_i la columna i de V , definimos para $i = 1, \dots, n$ la *transformación característica* del dígito x_i como el vector $\mathbf{tc}(x_i) = (v_1^t x_i, v_2^t x_i, \dots, v_\alpha^t x_i) \in \mathbb{R}^\alpha$, donde $\alpha \in \{1, \dots, m\}$ es un parámetro de la implementación. Este proceso corresponde a extraer las α primeras *componentes principales* de cada imagen. La intención es que $\mathbf{tc}(x_i)$ resuma la información más relevante de la imagen, descartando los detalles o las zonas que no aportan rasgos distintivos. Dada una nueva imagen x de un dígito, se calcula $\mathbf{tc}(x)$ y se compara con $\mathbf{tc}(x_i)$, para $i = 1, \dots, n$, utilizando algún criterio de clasificación adecuado, como por ejemplo kNN .

Finalmente, nos concentramos en la evaluación de los métodos. Dado que necesitamos conocer a qué dígito corresponde una imagen para poder verificar la correctitud de la clasificación, una alternativa es particionar la base de entrenamiento en dos, utilizando una parte de ella en forma completa para el entrenamiento y la restante como test, pudiendo así corroborar la predicción realizada. Sin embargo, realizar toda la experimentación sobre una única partición de la base podría resultar en una incorrecta estimación de parámetros, como por ejemplo el conocido *overfitting*. Luego, se considera la técnica de *cross validation*, en particular el *K-fold cross validation*, para realizar una estimación de los parámetros del modelo que resulte estadísticamente más robusta.

Enunciado

Se pide implementar un programa en C o C++ que lea desde archivos las imágenes de entrenamiento correspondientes a distintos dígitos y que, utilizando los métodos descriptos en la sección anterior, dada una nueva imagen de un dígito determine a qué número pertenece. Para ello, el programa deberá implementar el algoritmo de kNN así como también la reducción de dimensión previa utilizando PCA.

Con el objetivo de obtener la descomposición en valores singulares, se deberá implementar el método de la potencia con deflación para la estimación de autovalores/autovectores de la matriz de covarianza. En este contexto, la factibilidad de aplicar este método es particularmente sensible al tamaño de las imágenes de la base de datos. Por ejemplo, considerar imágenes en escala de grises de 100×100 píxeles implicaría trabajar con matrices de tamaño 10000×10000 . Se pide tener en cuenta este factor durante el desarrollo y analizar cómo afecta (si es que lo hace) en el desarrollo del trabajo.

Consideramos la base de datos MNIST, en la versión disponible en *Kaggle*. Esta base contiene un conjunto de datos de entrenamiento de 42.000 dígitos manuscritos en escala de grises y con $M = 28$. A su vez, provee un conjunto de datos de test de 28.000 dígitos, de similares características, pero sin el correspondiente etiquetado. En base a estos datos, se pide separar el procedimiento en dos etapas.

La primera de ellas consiste en realizar un estudio experimental con los métodos propuestos sobre la base de entrenamiento y utilizando la técnica de *K-fold cross validation* mencionada anteriormente. Para esta última tarea en particular, se recomienda leer y utilizar la rutina `cvpartition` provista por MATLAB. Llamamos k a la cantidad de vecinos a considerar en el algoritmo kNN , α a la cantidad de componentes principales a tomar y K a la cantidad de particiones consideradas para el cross-validation. La calidad de los resultados obtenidos serán analizados considerando la tasa de efectividad lograda, es decir, la cantidad dígitos correctamente clasificados respecto a la cantidad total de casos analizados. En función de la experimentación se piden como mínimo los siguientes experimentos:

- Analizar el comportamiento y la factibilidad de aplicar el algoritmo kNN para un rango razonable de valores distintos de k , analizando el compromiso entre el tiempo de ejecución y la calidad de los resultados obtenidos. Es posible, como experimento opcional, considerar alguna técnica alternativa de preprocesamiento de la información en caso de ser necesario (y, desde ya, que no sea PCA).

- Analizar la calidad de los resultados obtenidos al combinar PCA con kNN , para un rango amplio de combinaciones de valores de k y α . Considerar en el análisis también el tiempo de ejecución.
- Realizar los experimentos de los ítems anteriores para al menos dos valores distintos de K . Justificar el por qué de la elección de los mismos.
- En base a los resultados obtenidos para ambos métodos, seleccionar aquella combinación de parámetros que se considere la mejor alternativa, con su correspondiente justificación, y compararlas entre sí y sugerir un método para su utilización en la práctica.

Finalmente, y con los parámetros seleccionados en la fase experimental, ejecutar el método seleccionado sobre el conjunto de datos de test, utilizando como entrenamiento los 42.000 dígitos comprendidos en el conjunto de entrenamiento. Analizar el tiempo de cómputo requerido. Dado que la cátedra no posee la información sobre a qué dígito corresponde cada imagen (y la idea no es graficar uno por uno y obtenerlo a mano), se sugiere a cada grupo participar en la competencia *Digit Recognizer* actualmente activa en *Kaggle* realizando el/los envíos que consideren apropiados y reportar en el informe los resultados obtenidos.

Puntos opcionales (no obligatorios)

- Mostrar que si tenemos la descomposición $A = U\Sigma V^t$, V es la misma matriz que obtenemos al diagonalizar la matriz de covarianzas.
- Realizar experimentos utilizando dígitos manuscritos creados por el grupo, manteniendo el formato propuesto.¹ Reportar resultados y dificultades encontradas.
- Implementar alguna mejora al algoritmo de kNN .

Programa y formato de archivos

Se deberán entregar los archivos fuentes que contengan la resolución del trabajo práctico. El ejecutable tomará tres parámetros por línea de comando, que serán el nombre del archivo de entrada, el nombre del archivo de salida, y el método a ejecutar (0: kNN , 1: PCA + kNN).

Asumimos que la base de datos de imágenes de entrenamiento se llama `train.csv` y que la base de test `test.csv`, y que ambas siguen el formato establecido por la competencia. Para facilitar la experimentación, el archivo de entrada con la descripción del experimento tendrá la siguiente estructura:

- La primera línea contendrá el path a la(s) base(s) de datos, k , α y K .
- Luego, habrá K líneas de 42.000 valores, uno por cada muestra de la base de entrenamiento, donde un 1 indicará que esa muestra se considera parte del entrenamiento, y un 0 que se considera parte del test. Luego, de esta forma se pueden codificar las particiones realizadas por el *K-fold cross validation*.

El archivo de salida obligatorio tendrá para cada partición que figure en el archivo de entrada el vector solución con los α valores singulares de mayor magnitud, con una componente del mismo por línea. Además, el programa deberá generar un archivo, cuyo formato queda a criterio del grupo, indicando la tasa de reconocimiento obtenida para cada partición. Adicionalmente, se generará un archivo que concatene la extensión `.csv` al segundo valor recibido como parámetro del programa, que escribirá la predicción realizada sobre la base de test en el formato requerido por la competencia siguiendo el formato establecido por las reglas de la competencia.

Junto con el presente enunciado, se adjunta una serie de scripts hechos en `python` y un conjunto instancias de test que deberán ser utilizados para la compilación y un testeo básico de la implementación. Se recomienda leer el archivo `README.txt` con el detalle sobre su utilización.

Sobre la entrega

¹Notar que en la base original las figuras están rotadas y es blanco sobre negro, y no al revés.

- **FORMATO ELECTRÓNICO:** Jueves 14 de Mayo de 2015, hasta las 23:59 hs., enviando el trabajo (informe + código) a `metnum.lab@gmail.com`. El asunto del email debe comenzar con el texto [TP2] seguido de la lista de apellidos de los integrantes del grupo. Ejemplo: [TP2] Acevedo, Miranda, Montero
- **FORMATO FÍSICO:** Viernes 15 de Mayo de 2015, de 17:30 a 18:00 hs.

6.1. Apéndice B:

```
double ObtenerAutovalorMayor(Matriz<double> &A, Matriz<double>& v0){

    double lambda= normaInf(v0);
    double normalizar= norma2(v0);

    for(int i = 0; i < v0.filas(); i++){ v0[i][0] = v0[i][0]/normalizar;}

    Matriz<double> v1(v0.filas(),1);
    int i = 0;
    for(;i<3000 ;i++)
    {

        v1= A* v0;

        lambda= normaInf(v1)/normaInf(v0);

        normalizar = norma2(v1);

        for(int j = 0; j < v1.filas(); j++){ v1[j][0] = v1[j][0]/normalizar; }
        if(haypocadif(v0, v1)){break;}
        v0=v1;
    }

    return lambda;
}

vector<double> metodoPotencias(Matriz<double>& A, unsigned int alpha, Matriz<double>& P,
Matriz<double>& v){

    vector<double> autovalores;

    for(int i=0;i<alpha;i++)
    {
        double autovalori= ObtenerAutovalorMayor(A,v);

        autovalores.push_back(autovalori);

        for(int j=0;j<A.columnas();j++){ P[j][i]=v[j][0]; }

        Matriz<double> vt(1, v.filas());

        vt= v.traspuesta();

        Matriz<double> prod(v.filas(),v.filas());
        prod= v*vt*autovalori;

        A =A - prod;
    }

    return autovalores;
}
```