Modelos matemáticos y numéricos

Proyecto #1: Reconocimiento de patrones

Instructor: JL Morales
Alumno: Omar Pardo 130013

Introducción

Clasificar automáticamente dígitos manuscritos es un problema ampliamente difundido en el área de reconocimiento de patrones. La aplicación más popular consiste en automatizar la lectura de los códigos postales en piezas de correo.

En este proyecto, los dígitos consisten de imágenes de 20×20 pixeles en tonos de gris; para fines prácticos las imágenes se tratan como vectores en $\mathbb{R}^{400}$, obtenidos apilando las 20 columnas de una imagen. Para cada dígito d, $d \in \{0, 1, \ldots, 9\}$, tenemos una muestra de 500 elementos organizada como una matriz de 400×500.

Dado que el tamao de la muestra no es tan grande, una manera de resolver el problema anterior es mediante la Descomposición en Valores Singulares (DVS).

Fundamentos Matemáticos

Teorema 1 (Descomposición en valores singulares)**.** *Si A es una matriz real de $m \times n$, entonces existen matrices ortogonales*

$$U = [u_1, \ldots, u_m] \in \mathbb{R}^{m \times m}, \quad V = [v_1, \ldots, v_n] \in \mathbb{R}^{n \times n}$$

tales que

$$U^T AV = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}, \quad p = \min\{m, n\}$$

en donde $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$.

Prueba 1. *Consideremos la norma matricial inducida por la norma vectorial euclidiana; entonces es claro que existen 2 vectores unitarios $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ con la propiedad*

$$Ax = \sigma y, \quad \sigma = ||A||_2,$$

Esto se cumple debido a que la norma 2 euclidiana es una funcin lineal en un conjunto compacto (la esfera unitaria), por lo que el mximo se alcanza, segn el Teorema de BolzanoWeierstrass. Por lo tanto, el supremo que da lugar a la norma matricial, es un mximo.

Ahora podemos construir matrices ortogonales U, V a partir de los vectores x, y como sigue

$$V = [\, x \,|\, V_2 \,], \quad U = [\, y \,|\, U_2 \,].$$

De la construcción anterior, haciendo lgebra es fácil probar que

$$U^T AV = \begin{bmatrix} \sigma & w^T \\ 0 & B \end{bmatrix} = A_1$$

en donde $w \in \mathbb{R}^{n-1}$ *y* $B \in \mathbb{R}^{(m-1)\times(n-1)}$*; de la desigualdad*

$$\left|\left| A_1 \begin{bmatrix} \sigma \\ w \end{bmatrix} \right|\right|_2^2 \geq (\sigma^2 + w^T w)^2$$

podemos concluir que $||A_1||_2^2 \geq \sigma^2 + w^T w$*. Sin embargo* $\sigma^2 = ||A||_2^2 = ||A_1||_2^2$ *(por la descomposicin recursiva que coloca a* σ *como el mximo valor singular), por lo tanto* $w = 0$*. Es decir que* A_1 *tiene la estructura diagonal por bloques*

$$A_1 = \begin{bmatrix} \sigma & 0 \\ 0 & B \end{bmatrix}$$

Un argumento inductivo evidente completa la prueba del resultado. □

Propiedades

La DVS revela prácticamente toda la estructura algebraica de una matriz. Por ejemplo:

1. Los valores singulares de A son las longitudes de los semi ejes del hiperelipsoide

   $$H = \{Ax : ||x||_2 = 1\}$$

   definido por A.

2. Si definimos r como

   $$\sigma_1 \geq \cdots \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0,$$

   entonces

   (a) $\text{rango}(A) = r$

   (b) $K(A) = \text{gen}\,\{v_{r+1}, \ldots, v_n\}$

   (c) $R(A) = \text{gen}\,\{u_1, \ldots, u_r\}$

3. La matriz A se puede representar como una suma de matrices de rango 1 (desarrollo DVS)

   $$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T. \tag{1}$$

4. Diversas relaciones entre $||A||_\text{F}$ y $||A||_2$

   (a) $||A||_\text{F} = \sigma_1^2 + \cdots + \sigma_p^2, \quad p = \min\{m, n\}$

   (b) $||A||_2 = \sigma_1$

   (c) $\min\limits_{||x||_2=1} ||Ax||_2 = \sigma_n, \quad (m \geq n).$

Fundamentos Numéricos

La DVS es construible de forma numéricamente estable. El método más común es el de la *bidiagonalización* y consiste de 2 fases:

1. Reducir a A a una forma bidiagonal usando una serie de transformaciones ortogonales. Esta fase es determinística y el tiempo que tarda en correr depende solamente de las dimensiones de la matriz A.

2. Remover los elementos de la diagonal superior de la matriz resultante de la fase anterior usando también transformaciones ortogonales. Esta fase es iterativa, pero converge rápidamente.

Enfoquémonos en la primera fase. El primer paso crea dos matrices ortogonales U_1 y V_1, de tal forma que:

$$U_1 A = \begin{bmatrix} a'_{11} & a'_{12} & ... & a'_{1n} \\ 0 & & & \\ ... & & B' & \\ 0 & & & \end{bmatrix}$$

$$U_1 A V_1 = \begin{bmatrix} a''_{11} & a''_{12} & 0 & ... & 0 \\ 0 & & & & \\ ... & & B'' & & \\ 0 & & & & \end{bmatrix}$$

Si A es de $m \times n$, entonces B'' es de $(m-1) \times (n-1)$. El siguiente paso es trabajar de forma recursiva con B'' y seguir por ese camino, hasta producir las matrices U_1, ..., U_{n-1}, V_1, ..., V_{n-2}, de tal forma que $U_{n-1}...U_1 A V_1...V_{n-2}$ es bidiagonal (asumiendo que $m \geq n$).

En ambas fases las transformaciones ortogonales son construidas usando matrices de Householder. El costo computacional de este método es $\sim 2mn^2 + 11n^3$.

Metodología

Recordemos que para cada dígito d, $d \in \{0, 1, \ldots, 9\}$, tenemos una muestra de 500 elementos organizada como una matriz de 400×500.

Consideremos un dígito particular, digamos $d \in \{0, 1, \ldots, 9\}$. Si obtenemos la DVS de A^d, entonces la propiedad (1)

$$A^d = \sum_{i=1}^{r} \sigma_i u_i v_i^T.$$

indica que cada columna de A^d, denotada como a_j, se puede expresar como

$$a_j = \sum_{i=1}^{m} (\sigma_i v_{ij}) u_i,$$

en donde v_{ij} es la j-ésima componente del vector v_i. La expresión anterior sugiere que cada columna se puede aproximar por los primeros K términos de la suma, es decir

$$a_j \approx \sum_{i=1}^{K} (\sigma_i v_{ij}) u_i.$$

Notar que los primeros vectores singulares de A^d contienen la *información dominante.*

Ahora podemos identificar un dígito desconocido, digamos z, resolviendo el siguiente problema de mínimos cuadrados (MC).

$$\text{minimizar} \quad ||z - U_K \alpha||_2^2,$$

en donde U_K es la submatriz de U formada por los primeros K vectores singulares. La solución del problema de MC es

$$\alpha = U_K^T z,$$

mientras que la norma del residuo es

$$||r(z)||_2 = ||z - U_K U_K^T z||_2 \tag{2}$$

Es inmediato concluir que si $||r(z)||_2$ es muy pequeña, entonces probablemente $z = d$.

El razonamiento anterior se puede formular como el siguiente algoritmo:

1. Inicio.
   - Calcular la DVS de las matrices $A^j, \quad j = 0, \ldots, 9$.
2. Elegir z, un dígito de la muestra completa
   - resolver (1) para cada dígito
   - asignar z a la clase con menor error $||r(z)||_2$.
   - verificar si el dígito fue identificado correctamente.
3. Repetir 2. para todos los vectores z de la muestra y obtener la tasa de éxito de cada dígito.

Experimentos

La metodologa anterior se implementó en MATLAB usando el script *reconocimiento_digitos.m* y las funciones auxiliares *MUK.m* y *reconocimiento.m* (archivos adjuntos a este documento). Se realizaron predicciones de todos los vectores de la muestra para K de 1 a 5. Los porcentajes de acierto para cada dígito y cada K se encuentran en la *Table 1.*

Table 1: Porcentajes de acierto para cada dígito y distintas K's

dígito / K	1	2	3	4	5
0	0.918	0.936	0.968	0.978	0.984
1	0.946	0.986	0.988	0.986	0.978
2	0.768	0.858	0.902	0.912	0.924
3	0.806	0.844	0.882	0.894	0.914
4	0.784	0.804	0.882	0.896	0.902
5	0.644	0.806	0.852	0.902	0.898
6	0.886	0.94	0.962	0.964	0.974
7	0.832	0.868	0.908	0.924	0.926
8	0.756	0.83	0.868	0.882	0.91
9	0.768	0.872	0.912	0.922	0.908

Conclusiones

Es de sorprender que usando apenas 5 valores singulares de 500, logremos un clasificador que en el dígito que peor se comporta tiene prácticamente un 90% de acierto en sus predicciones. Es una gran muestra de la capacidad de la DVS. Independientemente de este innegable poder, no hay que olvidar que al tener una muestra pequea, toda fue usada para *entrenamiento* y no hubo de *prueba*, por lo que probablemente el clasificador este sobreajustado a la muestra. Es decir, si lo probamos con una nueva muestra posiblemente los resultados no serán tan buenos.

También resalta el hecho de que hay dígitos que consistentemente son clasificados correctamente con un porcentaje menor que los otros. Este es el caso del 5 o el 8. Mi hipótesis es que esto se debe a que son más propensos a ser confundidos con otros números, mientras que los que suelen tener un mayor porcentaje de clasificación correcta son en cierta medida distintos a todos los demás dígitos.