Compresión de datos aplicando la pseudoinversa de una matriz

Estudientes Grupo 2: Fabián Crawford Barquero,
Irene Muñoz Castro, Luis Morales Rodríguez,
Steven Badilla Soto, Adrián Trejos Salazar

I. Resumen

El valor singular de descomposición o SVD (Singular Value Decomposition) por sus siglas en inglés, es un método de factorización de matrices, el cual, descompone una matriz en tres más, trayendo una gran cantidad de propiedades que son muy útiles en la ciencia de datos. Una de las aplicaciones es la compresión de datos, especialmente en imágenes. Una dato que se pueda representar como una matriz, se puede descomponer en tres nuevas matrices que multiplicadas son equivalentes a la primera, y, esta descomposición se puede obtener utilizando la pseudoinversa de una matriz, o viceversa, con la SVD se puede obtener la inversa de una matriz.

II. Compresión de datos

Supondremos que tenemos una matriz que contiene datos en X1, X2,..., Xn, y, cada Xi es un vector, entonces, cada elemento Xi tendrá una característica o atributo. Como ejemplo, pensaremos que cada Xi es una persona, y que los valores en el vector son atributos relacionados con esa persona.

A. Problema

Tenemos una tabla de personas con sus nombres, la fecha de nacimiento y la cantidad de hijos que tuvieron en su vida, la tabla se presenta a continuación:

TABLE I: Información Cantidad Hijos

Tabla Cantidad de hijos		
Nombre	Año nacimiento	Cant. Hijos
Landen Jordan	1777	0
Cole McCoy	1838	12
Davion Powell	1752	0
Laurel Flood	1826	15
Angel Gray	1862	2
Grace Velez	1854	5
Elijah Rogers	1882	0
Lesly Wyatt	1815	0
Reid Williams	1835	2
Claudia Torres	1843	20

El objetivo es comprimir los datos de la columna 2 y 3.

B. Solución

Se utilizará un método también conocido como análisis del componente principal, el cual consigue convertir números grandes en unos más pequeños, con una correlación, o también, se puede utilizar para encontrar patrones en los datos.

Lo primero que se quiere hacer es encontrar un número promedio entre la distancia de los datos.

$$x - \bar{x} = (x_1 - \bar{x}, ..., x_n - \bar{x})\,(1)$$

Esto se consigue sumando todas las distancias y dividiéndolas entre la cantidad de datos menos uno -usar n-1 es necesario para convertir el resultado en un estimador imparcial, aumentando la precisión-. Como son dos vectores (Fecha de nacimiento y cantidad de hijos), se tiene que agregar la distancia de estos datos también.

$$y - \bar{y} = (y_1 - \bar{y}, ..., y_n - \bar{y})\,(2)$$

Al unir las funciones (1) y (2), y dividiéndolo por el n-1 mencionado anteriormente, obtenemos la covarianza, el cual mide la varianza entre x y y. La ecuación quedaría de la siguiente manera:

$$cov = \frac{\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{n - 1}\,(3)$$

Note que la ecuación (3) se puede representar como

$$cov(x, y) = \frac{(x - \bar{x})^T(y - \bar{y})}{n - 1}\,(4)$$

Por último podemos cambiaremos el valor centrado ($\bar{x}$) por un valor μ, talque:

$$\mu = \frac{1(x_1 + ... + x_n)}{n}\,(5)$$

Por lo tanto $\mu = (\mu_1 + ... + \mu_n)$. Estos valores se pueden encontrar utilizando la pseudoinversa.

$$\mu * A^+ = \bar{z}\,(6)$$

Donde μ es el vector de los valores que se quieren encontrar, A+ es la pseudoinversa de los valores de X y Y, y $\bar{z}$ es la solución al problema. Después de resolver y encontrar μ, obtenemos que $\mu_{n1} = 1828.4$ y $\mu_2 = 5.6$, aplicando la ecuación 4 sustituyendo μ_{n1} por y μ_2 por , para cada valor de la tabla, la compresión quedaría como:

TABLE II: Información Cantidad Hijos Info Comprimida

Tabla Cantidad de hijos Compresa		
Nombre	Año nacimiento	Cant. Hijos
Landen Jordan	-51.4	-5.6
Cole McCoy	9.6	6.4
Davion Powell	-76.4	-5.6
Laurel Flood	-2.4	9.4
Angel Gray	33.6	-3.6
Grace Velez	25.6	-0.6
Elijah Rogers	53.6	-5.6
Lesly Wyatt	13.4	-5.6
Reid Williams	6.6	-3.6
Claudia Torres	14.6	14.4

Para este problema, se puede encontrar el valor de μ de una forma más sencilla, ya que los valores son sencillos y fáciles de relacionar, y el objetivo de aplicar el método sobre una cantidad de datos tan pequeña es meramente ilustrativo, pero, cuando se trata de matrices con una cantidad de datos considerable o con valores más complejos como los de una imágen, se utiliza este método consiguiendo así, por ejemplo, comprimir los valores de una matriz.

III. Referencias

[1] https://www.cis.upenn.edu/ cis515/cis515-12-sl13.pdf