Análisis de Componentes Principales

Luis Escobar

15 September 2018

Brevísimo repaso

Para entender el procedimiento que se realiza con el método de componentes principales, vale la pena hacer un breve recuento de algunos conceptos y médidas estadísticas importantes: Normalización, Matriz de Covarianzas, Matriz de Correlaciones, Eigenvalores y Eigenvectores.

Normalización

Recordemos que podemos normalizar (i.e., transformar una serie de datos para que tenga media cero y varianza 1) restando a cada punto de la serie la media de los datos y escalando con la varianza. El escalamiento se usa principalmente cuando puede haber problemas de escalas en nuestros datos. Por ejemplo, si estamos realizando un análisis en el que tenemos una variable que mida el salario de una muestra de personas, y otra que mida su estatura, nos encontraremos con que, dado que el salario generalmente se encuentra en miles de pesos, y la estatura se encuentra en metros, en cualquier análisis que realicemos, los resultados estarán sumamente sesgados por la magnitud de la variable "salario", por lo que, para hacer comparables las dos variables, es una buena práctica normalizar los datos. Para normalizar nuestros datos, hacemos:

$$\frac{X - \overline{X}}{Var(X)} \tag{1}$$

Matriz de Covarianzas.

Mientras que la media y la varianza son medidas de tendencia central y de dispersión, respectivamente, para una serie de datos, existen medidas similares para la interacción entre dos o más series de datos. Una de esas medidas es la covarianza. La covarianza de dos series se calcula como:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{2}$$

La covarianza de una serie consigo mismo da como resultado su varianza. Cuando se tienen dos o más series, podemos acomodar las covarianzas en una matriz.

Matriz de Correlaciones

Definimos la correlación entre dos variables como:

$$corr_{xy} = \frac{s_{xy}}{Var(x)Var(y)} \tag{3}$$

Es decir, es el cociente de la covarianza entre el producto de sus varianzas. El coeficiente de correlación nos dice qué tan "sincronizados" son los movimientos entre las variables. Si la correlación es igual a 1 (valor máximo) quiere decir que el aumento (disminución) de una serie tiene un movimiento en la misma dirección y

de la misma magnitud en la otra serie, análogamente, si la correlación es cero el movimiento es en dirección contraria. Cuando tenemos dos o más series también es posible crear una matriz de correlaciones.

Eigenvalores

Sea A una matriz de tamaño $n \times n$ y sea I la matriz identidad del mismo tamaño. Los eigenvalores de A son los valores $\lambda_1, \lambda_2, \ldots, \lambda_n$ que satisfacen la ecuación polinomial $|A - \lambda I| = 0$ (donde $||$ es el determinante).

Referencias