Técnicas de visualización y reducción de dimensiones en espacios de alta dimensionalidad

Mtro. José Gustavo Fuentes Cabrera

Facultad de Estudios Superiores
Acatlán

Apuntes de Temas Selectos de Estadística

Licenciatura en Matemáticas Aplicadas y Computación

Índice

1. Introducción

En esta unidad, revisaremos lo correspondiente a las técnicas más utilizadas para analizar visualmente un espacio de alta dimensionalidad, así como reducir dimensiones para simplificar el problema estadístico planteado.

2. Análisis de Componentes Principales

Es un método estadístico multivariante que se clasifica como método de simplificación o de reducción de la dimensión. El ACP permite describir de forma sintética la estructura e interrelación de las variables originales. El método tiene por objeto transformar un conjunto de variables en un nuevo conjunto denominado componentes principales. Los nuevos componentes tienen la característica de ser incorrelacionados (ortogonales) y se ordenan de acuerdo a la cantidad de información (varianza) que llevan incorporada. Las componentes principales se expresan como una combinación lineal de las variables originales. Revisemos un esbozo de su obtención. Sean $X_1, X_2, \ldots, X_p$ un conjunto de variables de una muestra de tamaño n interrelacionadas entre sí, se busca obtener otro conjunto $Z_1, Z_2, \ldots, Z_k$ con $k \leq p$ tales que sean una combinación lineal del conjunto inicial y que expliquen la mayor parte de su variabilidad.

Obtengamos la primera componente:

$$Z_{1i} = u_{11}X_{1i} + u_{12}X_{2i} + \ldots + u_{1p}X_{pi}$$

Al tomar las n observaciones muestrales, tenemos:

$$\begin{bmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{1n} \end{bmatrix} = \begin{bmatrix} X_{11} & X_{21} & \ldots & X_{p1} \\ X_{12} & X_{22} & \ldots & X_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \ldots & X_{pn} \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{12} \\ \vdots \\ u_{1p} \end{bmatrix}$$

O en notación matricial:

$$\vec{Z_1} = X\vec{u}_1$$

Al suponer que las X_j están estandarizadas, podemos asumir que:

$$E[\vec{Z_1}] = E[X\vec{u}_1] = E[X]\vec{u}_1 = 0$$

Y la varianza sería:

$$\begin{aligned} V[\vec{Z_1}] &= \frac{1}{n}\sum_{i=1}^{n} Z_{1i}^2 \\ &= \frac{1}{n}\vec{Z_1}^T\vec{Z_1} \\ &= \frac{1}{n}\vec{u_1}^T X^T X \vec{u_1} \\ &= \vec{u_1}^T \left[\frac{1}{n}X^T X\right]\vec{u_1} \\ &= \vec{u_1}^T V \vec{u_1} \end{aligned}$$

Donde V es la matriz de covarianzas.
Para hallar $\vec{Z_1}$ necesitamos maximizar la varianza tal que la suma de los pesos u_{1j} al cuadrado sea igual a la unidad, en consecuencia tenemos un problema de optimización.

$$\begin{aligned} &max\ V[\vec{Z_1}] = \vec{u_1}^T V \vec{u_1} \\ &s.a.\ \sum_{j=1}^{p} u_{1i}^2 = \vec{u_1}^T\vec{u_1} = 1 \end{aligned}$$

Para resolverlo, recurrimos a los multiplicadores de Lagrange:

$$\begin{aligned} L &= \vec{u_1}^T V \vec{u_1} - \lambda\left(\vec{u_1}^T\vec{u_1} - 1\right) \\ \frac{\partial L}{\partial \vec{u_1}} &= 2V\vec{u_1} - 2\lambda\vec{u_1} = 0 \Rightarrow (V - \lambda I)\,\vec{u_1} = 0 \end{aligned}$$

La ecuación anterior solo tiene solución si $\|V - ambdaI\| = 0$ y en consecuencia, λ es un valor propio de la matriz V. Al premultiplicar por $\vec{u_1}^T$, tenemos:

$$\vec{u_1}^T (V - \lambda I)\,\vec{u_1} = 0 \Rightarrow \vec{u_1}^T V\vec{u_1} - \lambda\vec{u_1}^T I\vec{u_1} = \vec{u_1}^T V\vec{u_1} - \lambda = 0 \Rightarrow \vec{u_1}^T V\vec{u_1} = \lambda = V[\vec{Z_1}]$$

Sabemos que $\lambda_1, \lambda_2, \ldots, \lambda_n$ pueden ordenarse de forma ascendente tal que: $\lambda_1 > \lambda_2 > \ldots > \lambda_n$, de esta manera maximizaremos la varianza explicada tomando el mayor valor propio de V.

3. Escalamiento multidimensional

Otra técnica que es de utilidad para disminuir la dimensión de nuestra matriz entrada es conocida como escalamiento multidimensional(MDS por sus siglas en lengua inglesa). Este procedimiento consiste en generar un mapeo entre un espacio de alta dimensionalidad a otro de más baja dimensionalidad conservando la relación entre las distancias. En MDS clásico, las distancias se consideran euclídeas, considérese una matriz de $n \times p$, se genera una matriz de distancias y a partir de ella, se busca un mapeo en $\mathbb{R}^2$ o $\mathbb{R}^3$ tal que produzca con suficiente cercanía dicha matriz. Sea d_{ij} la distancias entre las observaciones $x^{(i)}$ y $x^{(j)}$ del espacio original, definimos δ_{ij} como la distancia correspondiente en dimensión reducida que producirá los vectores $x'^{(i)}$. Buscamos entonces un mecanismo que nos permita minimizar la discrepancia entre las distancias presentadas, naturalmente podemos proponer:

$$\sum_{i<j} \left(d_{ij} - \delta_{ij}\right)^2$$

Sin embargo, aunque matemáticamente correcto, presenta el incoveniente de ser sensible a la escala del espacio. Es por ello que incorporamos un factor normalizante a esta función de costo y adicionalmente convertimos a unidades lineales para tener:

$$\sqrt{\frac{\sum_{i<j} \left(d_{ij} - \delta_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}}$$

Que es conocida como *stress normalizado*. A continuación, mostramos los pasos necesarios para MDS clásico:

1. Generar la matriz de proximidades cuadradas $P^{(2)} = [p^2]$
2. Aplicar doble centrado $B = -\frac{1}{2} J P^{(2)} J$ usando la matriz $J = I - n^{-1} 1 1'$, donde n es el número de objetos.
3. Extraer los m valores propios mayores y sus correspondientes vectores propios $\vec{e_1}, \vec{e_2}, \ldots, \vec{e_m}$.
4. Se genera una configuración m-espacial de los n objetos derivados de la matriz de coordenadas $X = E_m \Lambda_m^{\frac{1}{2}}$,donde E_m es la matriz de

los m vectores propios y Λ_m es la matriz diagonal de los m valores propios de B.

El algoritmo anterior garantiza la minimización del stress cuando las distancias involucradas son euclideas.

4. Incrustación de vecinos estocásticos distribuída t

Conocida como t-SNE por sus siglas en inglés, es una técnica no lineal de reducción dimensional, posee un enfoque similar a MDS, sin embargo, el trasfondo del mapeo dimensional es distinto. t-SNE minimiza la divergencia entre dos distribuciones: la distribución que mide las similitudes por pares en el conjunto de entrenamiento y la distribución de las similitudes en el espacio de baja dimensionalidad. Considérese un conjunto de alta dimensionalidad $S_n = \{x^{(i)}, i = 1, 2, \ldots, n\}$ junto con una función de distancia $d\left(x^{(i)}, x^{(j)}\right)$. El objetivo es aprender una incrustación s-dimensional en el que cada objeto será representado por un punto $E = \{y^{(i)}, i = 1, 2, \ldots, n\}$ donde $y^{(i)} \in \mathbb{R}^s$, $s \in \{2, 3\}$ donde t-SNE definirá la probabilidad conjunta p_{ij} que mide la similitud por pares entre los vectores $x^{(i)}$ y $x^{(j)}$ al simetrizar dos probabilidades condicionales como sigue:

$$p_{j|i} = \frac{\exp\left(-d\left(x^{(i)}, x^{(j)}\right)^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-d\left(x^{(i)}, x^{(k)}\right)^2 / 2\sigma_i^2\right)}, \quad p_{i|i} = 0$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

La dispersión σ_i se elige de tal suerte que la perplejidad (medida de que tan bien una distribución de probabilidad predice una muestra, $2^{H(p)} = 2^{-\sum_x p(x) \log_2 p(x)}$) de la distribución condicional P_i es igual a una perplejidad definida u. En la incrustación s-dimensional E, las similitudes entre los vectores $y^{(i)}$ y $y^{(j)}$ se miden mediante un kernel de colas pesadas, en particular, la similitud incrustada q_{ij} para los vectores $y^{(i)}$ y $y^{(j)}$ se calcula mediante un kernel normalizado t-Student con un grado de libertad:

$$q_{ij} = \frac{\left(1 + \|y^{(i)} - y^{(j)}\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y^{(k)} - y^{(l)}\|^2\right)^{-1}} \quad q_{ii} = 0$$

Las colas pesadas del kernel normalizado t-Student permiten que vectores de entrada con disimilitud $x^{(i)}$ y $x^{(j)}$ sean modelados por sus contrapartes en baja dimensión $y^{(i)}$ y $y^{(j)}$ que están lejos entre sí. La ubicación de los puntos incrustados $y^{(i)}$ se determinan mediante la minimización de la divergencia de Kullback-Leibler entre las distribuciones conjuntas P y Q:

$$C\left(E\right) = KL\left(P||Q\right) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Dada la asimetría de la divergencia de Kullback-Leibler, la función objetivo se enfoca en modelar valores altos de p_{ij}(vectores similares) mediante valores altos de q_{ij} (puntos cercanos en el espacio incrustado). La función objetivo es no convexa en el incrustamiento E y se optimiza típicamente mediante gradiente descendiente.

$$\frac{\partial C}{\partial y^{(i)}} = 4 \sum_{j \neq i} \left(p_{ij} - q_{ij}\right) q_{ij} \sum_{k \neq l} \left(1 + \|y^{(k)} - y^{(l)}\|^2\right)^{-1}$$