Técnicas de reducción de la dimensionalidad

Minería de Datos - Grado en Matemáticas

DSLAB

2025-10-08

Reducción de la dimensión

- **Objetivo**: identificar un conjunto más pequeño de variables que capturen la mayor parte de la información esencial contenida en el total de las variables originales
- **Ventajas**:
  - Reducir la cantidad de información utilizada, especialmente útil cuando se trabaja con grandes conjuntos de datos
  - Eliminación de problemas de correlación entre variables –> elimina la redundancia en los datos y previene posibles distorsiones en los resultados del análisis
  - Posibilidad de visualizar los datos de manera sencilla (a veces en 2D) –> facilita la interpretación y la comunicación de resultados
- **Desventaja**: falta de explicabilidad cuando las nuevas variables son una combinación de las originales (e.g, PCA)

Técnicas

- **Análisis de componentes principales (PCA)**
- Escalado multidimensional (MDS)
- Análisis de correspondencias
- **Selección de variables (Feature Selection)**
- Autoencoders
- t-SNE (t-Distributed Stochastic Neighbor Embedding)

Selección de variables

- Seleccionar el subconjunto de variables que proporcionen la misma información que el total de variables (el rendimiento del modelo debe ser igual o superior)
- Tipos:
  - Filter
  - Wrapper
  - Embedded

Selección de variables: Filter

- Previos al entrenamiento de un modelo de ML
- Estudian la relación entre la variable objetivo y el resto de variables usando alguna medida de relevancia
- En función de dicha medida, proporcionan un ranking de variables
- Poco costosos computacionalmente
- Medidas de relevancia:
  - Correlación de Pearson
  - Tests estadísticos: T-test, Chi-cuadrado

Selección de variables: Wrapper

- Involucran modelos de ML
- Usan modelos de ML para evaluar la calidad de distintos subconjuntos de variables en función de su capacidad predictiva
- Tienen en cuenta la interacción entre las variables
- Costoso computacionalmente
- Ejemplos:
  - Forward selection
  - Backward elimination
  - Stepwise selection

Selección de variables: Embedded

- Involucran modelos de ML
- Realizan la selección y evaluación de las variables como parte del entrenamiento del modelo de ML
- Ejemplos: árboles de decisión, random forest, regresión lasso

Análisis de componentes principales

- Se debe a Pearson (1901) y a Hotteling (1933)
- **Objetivo**: representar la información recogida en el conjunto de datos original mediante un número menor de variables que son combinaciones lineales de las originales y que están incorreladas entre sí
- Herramienta exploratoria
- Información - Variabilidad (Varianza)

Dada una matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)^T$ formada por n elementos (filas) y p variables (columnas), las componentes principales son p nuevas variables $\mathbf{z_j}, j = 1, \dots, p$ construidas como una combinación lineal de las originales:

$$\mathbf{Z} = \mathbf{AX} = \mathbf{a}_1\mathbf{x}_1 + \cdots + \mathbf{a}_p\mathbf{x}_p$$

donde $\mathbf{A}^t\mathbf{A} = \mathbf{I}$, es decir, las nuevas variables $\mathbf{Z}$ están incorreladas entre sí. Es decir, obtener las componentes principales $\mathbf{Z}$ es realmente hacer una transformación ortogonal de las variables originales $\mathbf{X}$ para lograr nuevas variables incorreladas entre sí.

En este proceso se quiere maximizar la información, es decir, la variabilidad recogida, por tanto se buscan transformaciones que maximicen la varianza.

Cálculo de la primera componente

- La primera componente será $\mathbf{z}_1 = \mathbf{a}_1^t \mathbf{X}$, $\mathbf{a}_1$ es un vector de constantes
- Maximizar varianza, pero no por incrementar el valor de $\mathbf{a}$: $\mathbf{a}_1^t \mathbf{a}_1 = 1$
- El cálculo de $\mathbf{z}_1$ depende de $\mathbf{a}_1$. Así, buscamos $\mathbf{a}_1$ tal que $\mathbf{z}_1$ tenga máxima varianza y se cumpla que $\mathbf{a}_1^t \mathbf{a}_1 = 1$
- Máxima varianza –> maximizar $Var(\mathbf{z}_1) = \mathbf{a}_1^t \Sigma \mathbf{a}_1$, siendo Σ la matriz de varianzas covarianzas

Cálculo de la primera componente

- Multiplicadores de Lagrange para maximizar una función ($max\ Var(\mathbf{z}_1) = \mathbf{a}_1^t \Sigma \mathbf{a}_1$) sujeta a restricciones ($\mathbf{a}_1^t \mathbf{a}_1 = 1$):

$$L(\mathbf{a}_1) = \mathbf{a}_1^t \Sigma \mathbf{a}_1 - \lambda(\mathbf{a}_1^t \mathbf{a}_1 - 1)$$

Derivamos e igualamos a 0:

$$\frac{\partial L}{\partial \mathbf{a}_1} = 2\Sigma \mathbf{a}_1 - 2\lambda \mathbf{a}_1 = 0 \leftrightarrow \Sigma \mathbf{a}_1 = \lambda \mathbf{a}_1$$

Es decir, $\mathbf{a}_1$ es un autovector de la matriz Σ y λ su autovalor

Cálculo de la primera componente

- ¿Qué autovalor? Multiplicamos por $\mathbf{a}_1^t$: $\underbrace{\mathbf{a}_1^t \Sigma \mathbf{a}_1}_{Var(\mathbf{z}_1)} = \lambda \underbrace{\mathbf{a}_1^t \mathbf{a}_1}_{1} = \lambda$

  Es decir, λ es la varianza de la primera componente principal. Como se busca que sea máxima, será el mayor autovalor de Σ.

- La matriz de varianzas covarianzas Σ, de tamaño $(p \times p)$ es definida positiva: tiene p autovalores distintos $\lambda_1, \ldots, \lambda_p$

Segunda componente principal

Cálculo de la segunda componente principal $\mathbf{z}_2$: similar pero añadiendo la condición de estar incorrelada con $\mathbf{z}_1$ esto es,

$$Cov(\mathbf{z}_2, \mathbf{z}_1) = 0 \Leftrightarrow Cov(\mathbf{z}_2, \mathbf{z}_1) = Cov(\mathbf{a}_2^t\mathbf{X}, \mathbf{a}_1^t\mathbf{X}) = \mathbf{a}_2^t\Sigma\mathbf{a}_1 = 0$$

Como $\Sigma\mathbf{a}_1 = \lambda\mathbf{a}_1$: $\mathbf{a}_2^t\Sigma\mathbf{a}_1 = \lambda\mathbf{a}_2^t\mathbf{a}_1 = 0$ –> vectores ortogonales

Así, ahora, buscamos $max\ Var(\mathbf{z}_2) = \mathbf{a}_2^t\Sigma\mathbf{a}_2$ sujeta a $\mathbf{a}_2^t\mathbf{a}_2 = 1$ y $\mathbf{a}_2^t\mathbf{a}_1 = 0$

Y volvemos a llegar a $\Sigma\mathbf{a}_2 = \lambda\mathbf{a}_2$, escogiendo en este caso λ como el segundo mayor autovalor de la matriz de varianzas covarianzas

Finalmente obtenemos $\mathbf{z} = \mathbf{AX}$, siendo $\mathbf{z}_1, \ldots, \mathbf{z}_p$ variables incorreladas entre sí y con varianza

$$Var(\mathbf{z}_1) = \lambda_1, \ldots, Var(\mathbf{z}_p) = \lambda_p$$

que son los autovalores de la matriz de covarianzas.

Estos autovalores indican la variabilidad que recoge cada componente principal. Para hablar en términos de porcentaje de variabilidad recogido por cada componente:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

Objetivo: seleccionar el menor número $m < p$ de componentes principales que recojan un % alto de la variabilidad total

- El cálculo se basa en la matriz de varianzas covarianzas (no la de correlaciones) –> PCA depende de la escala
- PCA debe aplicarse a datos en los que las variables tengan escalas similares (comparables)

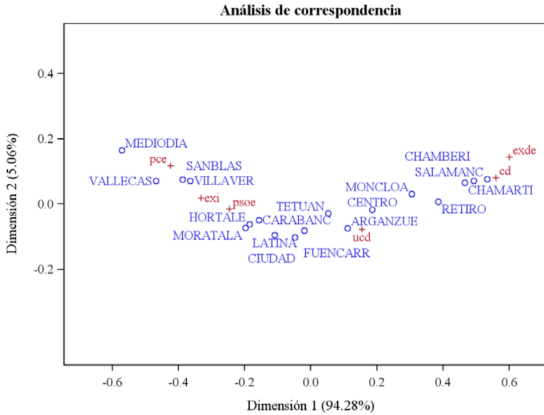
¿Qué ocurre si las variables originales están incorreladas?

Escalado multidimensional (MDS)

- Generalización del concepto de PCA
- Los datos son una matriz de distancias o de similitudes –> no se tienen observaciones y variables, se tienen distancias (o similitudes) entre ellos (Ej: comparación de productos para marketing)
- Traduce información de distancias entre n elementos en una representación de n puntos en el espacio más pequeño (2 o 3 dimensiones es lo ideal).
- La visualización en un espacio más pequeño permite entender la estructura, ver si hay grupos, qué elementos se asemejan más, etc.
- ¿Cómo funciona?
  - Ofrece unas coordenadas iniciales
  - Va modificando dichas coordenadas buscando que las distancias de las observaciones en las nuevas coordenadas sean lo más parecidas posibles a sus distancias en los datos originales.

Análisis de correspondencias

- Caso particular de MDS usando la distancia Chi-cuadrado
- Input: tabla de contingencia con las frecuencias absolutas observadas de 2 variables cualitativas. Ej: Color de ojos y color de pelo, Distritos de Madrid y partidos políticos



Referencias

John, George H, Ron Kohavi, y Karl Pfleger. 1994. «Irrelevant features and the subset selection problem». En *Machine learning proceedings 1994*, 121-29. Elsevier.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). *Feature extraction: foundations and applications* (Vol. 207). Springer.

Peña, D. (2013). *Análisis de datos multivariantes*. Cambridge: McGraw-Hill España.

Cuadras, C. M. (2007). *Nuevos métodos de análisis multivariante* (Vol. 20). Barcelona: CMC editions.