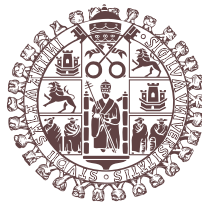


Universidad de Salamanca
Grado en Matemáticas

**REVISIÓN DE MÉTODOS
MULTIVARIANTES
SUPERVISADOS Y NO SUPERVISADOS**

Trabajo Fin de Grado



**VNiVERSiDAD
D SALAMANCA**

Alumno: Pedro Ángel Fraile Manzano

**Tutoras: Ana Belén Nieto Librero y Nerea González
García**

Salamanca, Julio de 2023

Índice general

Introducción	I
1 Métodos Supervisados	1
1.1 Teoría de decisión estadística	1
1.2 Redes Neuronales	2
2 Métodos no supervisados	3
2.1 Introducción	3
2.2 Análisis de Componentes Principales	4
2.2.1 Introducción	4
2.2.2 Definición y cálculo de las Componentes	4
2.2.3 Reducción de la dimensionalidad	7
Bibliografía	9

Introducción

El análisis multivariante se define como la rama del análisis estadístico que interpreta de manera simultánea la relación entre más de dos variables. Este rama del ha experimentado una gran expansión tanto en investigación como en aplicación debido al avance de la capacidad de computación de los actuales ordenadores haciendo que la posibilidad de

En la primera parte de la memoria se abordarán los llamados métodos supervisados. Estos son aquellos que, dado un conjunto de variables de entrada observadas $X_1 \dots X_p$ nos permiten predecir de distintas maneras una variable de salida. En este caso, el conjunto de datos recogidos y con los que se “entrena” al modelo contienen la observación de nuestra variable objetivo. Dentro de este tipo de métodos tendríamos métodos tan variados como los más simples métodos de regresión lineal multivariante, hasta las más complejas redes neuronales que podamos construir.

Por otro lado, los métodos no supervisados buscan relaciones entre las variables, de modo que no tenemos en el conjunto de entrenamiento ninguna información de cómo de correcto o incorrecto es lo que estamos afirmando. En este tipo de métodos entraría el apartado del análisis de componentes principales o

Capítulo 1

Métodos Supervisados

Supongamos que tenemos un conjunto de variables, \mathbf{X} que influyen sobre una o más variables conocidas \mathbf{Y} . A partir de ahora, las llamaremos variables de entrada y variables objetivo respectivamente. El principal propósito de los métodos supervisados, es dado una muestra de individuos con observaciones de ambos tipos de variables predecir la variable objetivo para nuevos individuos de los que solo conozcamos las variables de entrada

Para empezar tenemos que fundamentar de manera teórica como calcular esa función para predecir.

1.1. Teoría de decisión estadística

Sea $X \in \mathbb{R}^p$ un vector aleatorio real e $Y \in \mathbb{R}$ una variable aleatoria real. En este contexto, X e Y serán las variables de entrada y la variable de salida respectivamente. Asimismo, sea $\mathbb{P}(X, Y)$ la distribución de probabilidad conjunta.

Se busca una función $f(X)$ para predecir Y . Dicho predictor tiene asociada una pérdida, es decir, una forma de penalizar el error de predicción. En esta memoria, a no ser que se especifique utilizaremos el error cuadrático para las regresiones, $L(Y, f(X)) = (Y - f(X))^2$.

Definición 1.1.1. Llamaremos error de predicción esperado de f o $EPE(f)$ a la siguiente expresión:

$$EPE(f) = E(Y - f(X))^2 = \int (y - f(x))^2 \mathbb{P}(dx, dy) \quad (1.1)$$

A priori se conocen los valores de X , entonces si condicionamos a dichos valores, obtenemos que $\mathbb{P}(Y, X) = \mathbb{P}(Y|X) \cdot \mathbb{P}(X)$ aplicándolo en la expresión anterior resulta que

$$\begin{aligned} EPE(f) &= \int (y - f(x))^2 \mathbb{P}(dx, dy) = \int \int (y - f(x))^2 \mathbb{P}(dy|dx) \mathbb{P}(dx) \\ &= \mathbb{E}_X(\mathbb{E}_{Y|X}((Y - f(X))^2|X)) \end{aligned} \quad (1.2)$$

Este parámetro nos ofrece un criterio para encontrar f , es decir, f será la que minimice el $EPE(f)$, en concreto, $f(x) = \mathbb{E}(Y|X = x)$. Añadir que este sería el caso de la regresión.

Sea ahora G una variable categórica con K categorías posibles y \mathcal{G} el conjunto de categorías posibles. Se define la matriz de pérdida \mathbf{L} de tamaño $K \times K$ en el que el término $l_{i,j}$ es la pérdida que se da al clasificar G_i como G_j . De manera habitual se toma la pérdida 0 – 1 que se define como $l_{i,j} = 1 - \delta_{ij}$.

De esta manera, la matriz de pérdida es *simétrica y no negativa*.

Con esta función de pérdida el error de predicción esperado pasa a ser:

$$EPE(f) = \mathbb{E}(L(G, \hat{G})) = \mathbb{E}_X \sum_{k=1}^K (L(\mathcal{G}_k, \hat{G}(X)) \mathbb{P}(\mathcal{G}_k|X)) \quad (1.3)$$

De esta ecuación resulta que el predictor \hat{G} se explicita de la siguiente manera en el caso de la pérdida 0 – 1.

$$\hat{G}(X) = \mathcal{G}_k \quad \text{si} \quad \mathbb{P}(\mathcal{G}_k|X = x) = \max_{g \in \mathcal{G}} \mathbb{P}(g|X = x) \quad (1.4)$$

1.2. Redes Neuronales

Las redes neuronales artificiales, redes neuronales simplemente a partir de ahora, se basan en el funcionamiento básico de las neuronas biológicas. Este tipo de células reconocen

Capítulo 2

Métodos no supervisados

2.1. Introducción

2.2. Análisis de Componentes Principales

2.2.1. Introducción

2.2.2. Definición y cálculo de las Componentes

Sea un vector aleatorio $\mathbf{X}^T = [X_1, \dots, X_p]$ con vector de medias μ y matriz de covarianzas Σ .

Definición 2.2.1. Las componentes principales son combinaciones lineales de las variables $X_1 \dots X_p$

$$\mathbf{Y}_j = a_{1j}X_1 + \dots a_{pj}X_p = \mathbf{a}_j^T \mathbf{X} \quad (2.1)$$

Donde \mathbf{a}_j es un vector de constantes y la variable \mathbf{Y}_j cumple lo siguiente:

- Si $j = 1$ $Var(\mathbf{Y}_1)$ es máxima restringido a $\mathbf{a}_1^T \mathbf{a}_1 = 1$
- Si $j > 1$ debe cumplir:
 - $Cov(\mathbf{Y}_j, \mathbf{Y}_i) = 0 \quad \forall i < j$
 - $\mathbf{a}_j^T \mathbf{a}_j = 1$
 - $Var(\mathbf{Y}_j)$ es máxima.

De esta manera, estamos buscando una nueva base que consiga reunir las direcciones de máxima variación.

El cálculo de la primera componente principal se lleva a cabo con un proceso de optimización de la función $Var(\mathbf{Y}_1)$ sujeto a la restricción de que $\mathbf{a}_1^T \mathbf{a}_1 = 1$.

Aplicando el método de los multiplicadores de Lagrange, dada una función $f(\mathbf{x}) = f(x_1, \dots, x_p)$ diferenciable con una restricción $g(\mathbf{x}) = g(x_1, \dots, x_p) = c$ entonces existe una constante λ de manera la ecuación:

$$\frac{\partial f}{\partial x_i} - \lambda \frac{\partial g}{\partial x_i} = 0 \quad i = 1, \dots, p \quad (2.2)$$

Tiene como solución los puntos estacionarios de $f(\mathbf{x})$. Entonces, se puede definir la función $L(\mathbf{x}) = f(\mathbf{x}) - \lambda[g(\mathbf{x}) - c]$ que permite simplificar la expresión anterior a:

$$\frac{\partial L}{\partial \mathbf{x}} = 0 \quad (2.3)$$

Para el caso de las componentes principales, la función objetivo es la varianza de la combinación lineal, es decir, $f(\mathbf{x}) = \mathbf{x}^T \Sigma \mathbf{x}$ y la restricción aplicada es $g(\mathbf{x}) = \mathbf{x}^T \mathbf{x} = 1$.

Tomando $\mathbf{x} = \mathbf{a}_1$ se puede establecer $L(\mathbf{a}_1) = \mathbf{a}_1^T \Sigma \mathbf{a}_1 - \lambda[\mathbf{a}_1^T \mathbf{a}_1 - 1]$. Que al derivarla se obtiene:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{a}_1} &= 2\Sigma \mathbf{a}_1 - 2\lambda \mathbf{a}_1 \\ &= 2(\Sigma - \lambda) \mathbf{a}_1 \end{aligned}$$

Igualando a 0 tenemos la siguiente ecuación:

$$(\Sigma - \lambda I) \mathbf{a}_1 = 0 \quad (2.4)$$

Para que \mathbf{a}_1 sea un vector no trivial, tenemos que elegir λ de tal manera que $|\Sigma - \lambda I| = 0$, es decir, λ es un vector propio de la matriz de covarianzas, Σ . Al ser ésta una matriz semidefinido positiva y simétrica, los valores propios son reales y positivos. Por tanto, \mathbf{a}_1 es un vector propio de la matriz de covarianza.

La función a maximizar es $Var(\mathbf{Y}_1) = Var(\mathbf{a}_1^T \mathbf{X}) = \mathbf{a}_1^T \Sigma \mathbf{a}_1 = \mathbf{a}_1^T \lambda \mathbf{a}_1$, y para maximizarla basta tomar $\lambda = \max\{\lambda_1 \dots \lambda_p\}$. reordenando si es necesario, se tiene que $\lambda = \lambda_1$

Una vez calculada la primera componente principal \mathbf{Y}_1 , la segunda componente se calcula de manera análoga, maximizando $Var(\mathbf{Y}_2) = Var(\mathbf{a}_2^T \mathbf{X})$ condicionada por $\mathbf{a}_2^T \mathbf{a}_2 = 1$. A esta restricción tenemos que añadir la restricción $Cov(\mathbf{Y}_1, \mathbf{Y}_2) = 0$

Proposición 2.2.1. La condición $Cov(\mathbf{Y}_1, \mathbf{Y}_2) = 0$ equivale a la condición $\mathbf{a}_2^T \mathbf{a}_1 = 0$.

Demostración. Utilizando que $\mathbf{Y}_j = \mathbf{a}_j^T \mathbf{X} \quad \forall j$, tenemos entonces que:

$$\begin{aligned} Cov(\mathbf{Y}_2, \mathbf{Y}_1) &= Cov(\mathbf{a}_2^T \mathbf{X}, \mathbf{a}_1^T \mathbf{X}) \\ &= \mathbb{E}(\mathbf{a}_2^T (\mathbf{X} - \mu)(\mathbf{X} - \mu)^T \mathbf{a}_1) \\ &= \mathbf{a}_2^T \mathbb{E}((\mathbf{X} - \mu)(\mathbf{X} - \mu)^T) \mathbf{a}_1 \\ &= \mathbf{a}_2^T \Sigma \mathbf{a}_1 \\ &= \mathbf{a}_2^T \lambda_1 \mathbf{a}_1 \end{aligned}$$

De manera que, si $\mathbf{a}_2^T \lambda_1 \mathbf{a}_1 = 0 \Rightarrow \mathbf{a}_2^T \mathbf{a}_1 = 0$, luego son vectores ortogonales entre sí. \square

Observación: Esta proposición se puede extender de manera simple al caso de tener que calcular la i -ésima componente principal habiendo calculado las anteriores de las cuales se sepan los valores propios asociados.

Corolario 2.2.1. Las componentes principales son todas ortogonales entre sí.

Tomando la matriz formada por los p vectores propios como columnas tenemos la matriz ortogonal $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_p]$, de manera que el vector aleatorio

$$\mathbf{Y} = [Y_1, \dots, Y_p]^T = \mathbf{A}\mathbf{X}$$

Se deduce utilizando la ortogonalidad de $\mathbf{A} \Rightarrow \mathbf{A}^T \mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$ que:

$$\begin{aligned} Var(\mathbf{Y}) &= Var(\mathbf{A}\mathbf{X}) \\ &= \mathbb{E}((\mathbf{A}\mathbf{X})^T (\mathbf{A}\mathbf{X})) \\ &= \mathbb{E}(\mathbf{X}^T \mathbf{A}^T \mathbf{A} \mathbf{X}) \\ &= \mathbb{E}(\mathbf{X}^T \mathbf{X}) \\ &= Var(\mathbf{X}) \end{aligned}$$

Por tanto, se puede afirmar que las componentes principales contienen toda la variación del vector \mathbf{X} aleatorio inicial.

Sea ahora la matriz \mathbf{X} de datos de tamaño $n \times p$ resultante de tomar n observaciones de las p variables del vector aleatorio. Se supone además que esta es centrada y dividida todos sus elementos por $\sqrt{n-1}$.

Dado este marco la matriz \mathbf{X} tiene una descomposición en valores singulares $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$. Donde tanto \mathbf{U} como \mathbf{V} son matrices ortogonales. Esto implica que $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$.

En particular, \mathbf{U} es la matriz de vectores singulares por la izquierda de tamaño $n \times r$, \mathbf{V} la matriz de vectores singulares por la derecha de tamaño $p \times p$ y la matriz Σ es la matriz diagonal de valores singulares generalmente ordenados de forma decreciente, es decir Σ^2 es la matriz diagonal que contiene todos los valores propios de las matrices $\mathbf{X}^T\mathbf{X}$ y $\mathbf{X}\mathbf{X}^T$.

Por las transformaciones previas realizadas la matriz de estimaciones de las covarianzas es $\hat{S} = \mathbf{X}^T\mathbf{X}$ en consecuencia, el proceso que se describía en primera instancia para un vector aleatorio en el que conocíamos su matriz de covarianzas se puede aplicar de manera análoga a la matriz de datos.

Esto implica que la obtención de las componentes principales en el caso de ser calculadas desde la matriz de datos también se obtienen calculando los vectores y valores propios de la matriz de covarianzas muestral en este caso.

Proposición 2.2.2. La matriz \mathbf{V} de tamaño $(p \times p)$ es la matriz que contiene los vectores para hacer la combinación lineal que definen las componentes principales.

Demostración. La matriz $\mathbf{X}^T\mathbf{X}$ es la matriz de covarianzas $(p \times p)$ por la descomposición en valores singulares tenemos que:

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= (\mathbf{U}\Sigma\mathbf{V}^T)^T(\mathbf{U}\Sigma\mathbf{V}^T) \\ &= \mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T \\ &= \mathbf{V}\Sigma^T\Sigma\mathbf{V}^T \\ &= \mathbf{V}\Sigma^2\mathbf{V}^T\end{aligned}$$

Que es la diagonalización de la matriz $\mathbf{X}^T\mathbf{X}$ donde los elementos de la diagonal de Σ son las raíces cuadradas de los valores propios de $\mathbf{X}^T\mathbf{X}$. A esto también hay que añadir que \mathbf{V} es la matriz que tiene como columnas los vectores propios, es decir, es la matriz con los vectores que forman las componentes principales. \square

2.2.3. Reducción de la dimensionalidad

La utilidad de esta técnica es reducir el tamaño de la matriz de datos para poder representar las observaciones como puntos de una subvariedad de dimensión $m < p$. Con este fin en mente, se deben definir los siguientes conceptos.

Definición 2.2.2. Sea $\mathbf{A} \in \mathbb{M}_{n \times m}(\mathbb{R})$ definimos la *norma de Frobenius* de la matriz \mathbf{A} como :

$$\|\mathbf{A}\|_F = (tr(\mathbf{A}^T \cdot \mathbf{A}))^{\frac{1}{2}} = \left(\sum_{i=1}^n \sum_{j=1}^m a_{ij}^2 \right)^{\frac{1}{2}} \quad (2.5)$$

Proposición 2.2.3. La norma de Frobenius es invariante a transformaciones ortogonales

Demostración. Sea \mathbf{U} una matriz ortogonal, que cumple $\mathbf{U}^T \cdot \mathbf{U} = \mathbf{U} \cdot \mathbf{U}^T = I$, sea una matriz cualquiera \mathbf{A} , entonces:

$$\begin{aligned} \|\mathbf{U} \cdot \mathbf{A}\|_F^2 &= tr((\mathbf{U}\mathbf{A})^T \cdot (\mathbf{U}\mathbf{A})) \\ &= tr((\mathbf{A}^T \mathbf{U}^T) \cdot \mathbf{U}\mathbf{A}) \\ &= tr(\mathbf{A}^T \mathbf{A}) \\ &= \|\mathbf{A}\|_F^2 \end{aligned} \quad (2.7)$$

□

Aunque hay más normas que son invariantes ante transformaciones ortogonales, nos centraremos en la norma de Frobenius.

Definición 2.2.3. Dada una matriz $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$ de tamaño $n \times m$ y de rango r , entonces para $l < r$ se define la matriz truncada $\mathbf{A}_l = \mathbf{U}_l \Sigma_l \mathbf{V}_l^T$ donde:

- $\mathbf{U} := [\mathbf{U}_l, \mathbf{U}_{r-l}]$

-

-

Teorema 2.2.1 (De Eckart-Young). Sea una matriz \mathbf{A} de rango r , y sea una matriz \mathbf{B} rango $l < r$, entonces:

$$\|\mathbf{A} - \mathbf{B}\|_F \leq \|\mathbf{A} - \mathbf{A}_l\|_F \quad (2.6)$$

Proposición 2.2.4. Dada una matriz $\mathbf{A} =$

Bibliografía

- [1] **Chatfield,C y Collins A.J (1989).** *Introduction to multivariate analysis*, Chapman and Hall.
- [2] **Jolliffe I.T.(1986).** *Principal Component Analysis*, Springer-Verlag.
- [3] **Hastie, T.,Tibshirani, R. y Friedman J. (2001),** *The Elements of Statistical Learning, Data Mining, Inference and Prediction* Springer
- [4] **Cuadras, C.M. (2014),** *Nuevos métodos de Análisis Multivariante*, CMC Editions, Barcelona.