Universidad de Salamanca
Grado en Matemáticas

REVISIÓN DE MÉTODOS MULTIVARIANTES SUPERVISADOS Y NO SUPERVISADOS

Trabajo Fin de Grado

Alumno: Pedro Ángel Fraile Manzano

Tutoras: Ana Belén Nieto Librero y Nerea González García

Salamanca, Julio de 2023

Índice general

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 explicite 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:

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

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

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

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:

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

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

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

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 recogen

Capítulo 2

Análisis de Componentes Principales

2.1. Introducción

2.2. Definición y cálculo de las Componentes

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

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

$$\mathbf{Y}_j = a_{1j}X_1 + \ldots a_{pj}X_p = \mathbf{a}_j^T\mathbf{X} \tag{2.1}$$

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 $f(\mathbf{x}) = f(x_1 \ldots x_p)$ diferenciable con una restricción $g(\mathbf{x}) = g(x_1 \ldots x_p) = c$ entonces existe una constante λ de manera la ecuación:

$$\frac{\partial f}{\partial x_i} - \lambda \frac{\partial g}{\partial x_i} = 0 \quad i = 1, \ldots p \tag{2.2}$$

Tiene como solución los puntos estacionarios de $f(\mathbf{x})$

Sea ahora la función $L(\mathbf{x}) = f(\mathbf{x}) - \lambda[g(x) - c]$ entonces podemos simplificar la expresión anterior a:

$$\frac{\partial L}{\partial \mathbf{x}} = 0 \tag{2.3}$$

Para el cálculo de las componentes principales nuestra función a maximizar es la varianza de nuestra combinación lineal $\mathbf{Y}_1$

Para el cálculo de la $L(\mathbf{a}_1) = \mathbf{a}_1^T\Sigma\mathbf{a}_1 - \lambda[\mathbf{a}_1^T\mathbf{a}_1 - 1]$. Al derivarla obtenemos que :

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

Igualando a 0 tenemos la siguiente ecuación:

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

Para que la ecuación tenga una solución que no sea la trivial, tenemos que elegir λ de manera que $|\Sigma - \lambda I| = 0$. Luego λ es uno de los valores propios de la matriz. Generalmente una matriz $(p \times p)$ tiene p valores propios $\{\lambda_1, \ldots, \lambda_p\}$ y como $Var(\mathbf{Y}_1) = Var(\mathbf{a}_1^T\mathbf{X}) = \mathbf{a}_1^T\Sigma\mathbf{a}_1 = \mathbf{a}_1^T\lambda\mathbf{a}_1 = \lambda$ que es la variable a maximizar, elegimos $\lambda = \max\{\lambda_1, \ldots, \lambda_p\}$, por tanto, el vector $\mathbf{a}_1$ es el vector propio con valor propio $\lambda = \lambda_1$ reordenando si es necesario.

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:

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

De manera que, si $a_2^T\lambda_1 a_1 = 0 \Rightarrow a_2^T a_1 = 0$, luego son vectores ortogonales entre sí. □

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 sepamos sus valores propios.

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 $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]$,de manera que el vector aleatorio

$$\mathbf{Y} = [\mathbf{Y}_1, \ldots, Y_p]^T = \mathbf{AX}$$

De esta manera también obtenemos la diagonalización de la matriz de covarianzas

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \tag{2.5}$$

Y también podemos deducir que

2.3. Reducción de la dimensionalidad

Antes de continuar desarrollando, debemos definir unos conceptos previos.

Definición 2.3.1. Sea $\mathbf{A} \in \mathbb{M}_{n\times m}(\mathbb{R})$ definimos la *norma de Fröbenius* de la matriz $\mathbf{A}$ como :

$$||A||_F = (tr(A^T \cdot A))^{\frac{1}{2}} \tag{2.6}$$

Proposición 2.3.1. La norma de Fröbenius es invariante a transformaciones ortogonales

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

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

□

Corolario 2.3.1. Sea A una matriz y $A = U\Sigma V^T$ su descomposición en valores singulares, entonces $||A||_F = ||\Sigma||_F = \Sigma_{i=1}^n \sigma_i$ donde $\sigma_i, i = 1 \dots n$ son los valores singulares de A.

Teorema 2.3.1 (Teorema de Eckart-Young)**.**

Bibliografía

[1] **Chatfield,C y Collins A.J (1989)**. *Introduction to multivariate analysis*, Chapman and Hall.

[2] **Jollife 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.