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

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.

No obstante, para los casos de **clasificación** cambia el hecho de que la función cuadrática de pérdida no es adecuada. Sea G una variable categórica con K distintos valores posibles, entonces su función de pérdida se puede expresar como una matriz $\mathbf{L}$. Podemos definir el termino $l_{i,j}$ =“Pérdida de clasificar como G_i lo que en realidad es G_j”. Lo más frecuente es tomar la pérdida $0-1$, esta asigna 0 a las muestras correctamente clasificadas y 1 a las no categorizadas de manera satisfactoria.

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