Revisión de métodos multivariantes supervisados y no supervisados

Pedro Ángel Fraile Manzano

12 de julio de 2023

VNiVERSiDAD
D SALAMANCA

Contenido

Contenidos

Introducción

Definición

Los métodos multivariantes son el conjunto de técnicas que buscan describir y extraer información de las relaciones entre variables medidas en una o varias muestras u observaciones.

Introducción

Los métodos multivariantes pueden clasificarse de la siguiente manera según el caso de estudio:

- **Métodos supervisados:** estudian la relación estocástica que hay entre dos conjuntos de variables, un conjunto de variables predictoras y otro de variables respuesta.
  - *Regresión:* la o las variables variables respuesta son continuas.
  - *Clasificación:* la o las variables respuesta son cualitativas.
- **Métodos no supervisados:** analizan la estructura y relaciones que hay entre las variables.

Introducción

Objetivo:
Describir las principales técnicas multivariantes y dar un ejemplo de aplicación sobre datos reales de las mismas.

Contenidos

Métodos supervisados

Definición

Son aquellos que buscan inferir una relación estocástica entre dos grupos de variables, predictoras a partir de un conjunto de datos procedentes de observaciones simultáneas de ambos conjuntos.

Si se estudian varias variables de manera simultánea, se dividen en los siguientes tipos:

- Variables predictoras, son las variables independientes, se denotarán por el vector aleatorio $\mathbf{x} = [X_1, \ldots, X_p]$
- Variables respuesta, son las variables dependientes, se denotarán como el vector aleatorio $\mathbf{y} = [Y_1, \ldots, Y_K]$ y como Y cuando $K = 1$.

Métodos supervisados

Se trabajará con muestras aleatorias simples de N observaciones de las $K + p$ variables, obteniéndose las siguientes matrices:

Definición

Se llama matriz de datos $\mathbf{X}$ a la matriz de tamaño $N \times p$ cuyas filas $\mathbf{x}_i, \quad i = 1, \ldots, N$ representan una observación de las p variables predictoras.

Definición

Se llama matriz de respuestas $\mathbf{Y}$ a la matriz de tamaño $N \times K$ cuyas filas $\mathbf{y}_i, \quad i = 1, \ldots, N$ representan una observación de las K variables respuesta.

Métodos supervisados

Por tanto, se recogen N observaciones $(\mathbf{x}_i, \mathbf{y}_i)$ obteniéndose un vector de longitud $K + p$.
Los métodos supervisados buscan estimar una relación estocástica que se llamará predictor que conocido el valor de las variables predictoras $\mathbf{x}_0$, se pueda hacer una predicción del valor de las variables respuesta $\mathbf{y}_0$, que se denota por $\hat{\mathbf{y}}_0$.

Contenidos

Métodos supervisados

El objetivo de la regresión lineal es estudiar como un conjunto de variables respuesta están relacionadas con una combinación lineal de las variables predictoras.

Métodos supervisados

Tomamos el modelo en el que:

$$Y = f(\mathbf{x}) + \varepsilon \tag{2.1}$$

donde ε cumple:

- $\mathbb{E}(\varepsilon) = 0$
- $Var(\varepsilon) = \sigma^2$

A mayores se puede asumir que $\varepsilon \sim N(0, \sigma^2)$

Métodos supervisados

En la regresión lineal se asume que f es de la siguiente manera:

$$f(\mathbf{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{2.2}$$

- $\beta = [\beta_0, \beta_1, \ldots, \beta_p]^T$ es el vector con los parámetros de regresión.
- Añadiendo a $\mathbf{x}$ la variable $X_0 = 1$ se puede expresar f:

$$f(\mathbf{x}) = \mathbf{x}\beta \tag{2.3}$$

Tomando N observaciones, se puede obtener el vector $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = \mathbf{X}\beta \tag{2.4}$$

Se pueden estimar los parámetros de regresión mediante el método de los mínimos cuadrados:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.5}$$

Sustituyendo en la ecuación 2.4

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.6}$$

Métodos supervisados

Supóngase lo siguiente:

- $\mathbf{x} \sim N(\mu, \Sigma)$ del que obtenemos la matriz de datos $\mathbf{X}$ de tamaño $N \times p$.
- $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ donde $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_N)$.

Se cumplen las siguientes propiedades:

- El vector $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_N)$
- El vector $\hat{\beta}$ es un estimador insesgado.
- La varianza de $Var(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.
- El estimador $\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ es un estimador insesgado de σ^2 y además $\hat{\sigma}^2 \sim \frac{\sigma^2}{N-p-1}\chi^2_{N-p-1}$

Se puede definir un estadígrafo de contraste para comprobar si un conjunto de variables es estadísticamente significativo en el modelo.

$$F = \frac{\frac{(RSS_0 - RSS_1)}{p_1 - p_0}}{\frac{RSS_1}{N - p_1 - 1}} \sim F_{(p_1, p_0),(N-p_1-1)} \tag{2.7}$$

En el caso de que tengamos K variables respuesta $Y_1, \ldots, Y_K$ entonces el modelo es:

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \varepsilon_k = f_k(\mathbf{x}) + \varepsilon_k, \quad k = 1, \ldots K \tag{2.8}$$

donde $\varepsilon_k \sim N(0, \sigma_k^2)$ $\forall k = 1, \ldots, K$ que pueden o no estar correlacionados. Al tomar N observaciones se puede expresar de manera matricial

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{2.9}$$

donde $\mathbf{E}$ es la matriz $N \times K$ de errores cometidos en cada una de las observaciones.

El método de los mínimos cuadrados, cambia en el caso de que los errores tengan matriz de covarianzas $\mathbf{\Sigma}$ conocida el RSS se define de la siguiente manera:

$$RSS(\mathbf{B}, \mathbf{\Sigma}) = \sum_{i=1}^{N} (\mathbf{y}_i - f(\mathbf{x}_i)) \mathbf{\Sigma}^{-1} (\mathbf{y}_i - f(\mathbf{x}_i))^T \tag{2.10}$$

donde $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j) \mathbf{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{x}_j)^T}$ es la distancia de Mahalanobis entre dos observaciones.

Se puede utilizar el estadígrafo F de la ecuación 2.7 para seleccionar las variables del modelo.
También se puede definir lo siguiente:

Definición

Llamamos errores cuadrados acumulados penalizados, *PRSS*, a la suma de los cuadrados de los errores cometidos con el modelo lineal que usa el vector de parámetros β, añadiendo un término regulador $\lambda > 0$:

$$PRSS(\beta) = \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \qquad (2.11)$$

Esto permite dar una forma de regular los tamaños de los parámetros β

Contenidos

Métodos supervisados

La clasificación es el caso donde la variable respuesta es cualitativa. Hay dos tareas a realizar:

- Discriminación: Estudiar las características que diferencian a cada una de las poblaciones.
- Clasificación: Asignar nuevas observaciones a una de las poblaciones

Si tenemos dos poblaciones π_1, π_2 de las cuales se conocen las funciones de probabilidad de cada una de ellas f_1, f_2 , entonces si se conoce la probabilidad de de clasificación de clasificar en la población incorrecta, $P(1|2), P(2|1)$, entonces:

$$\begin{aligned} P(i|\mathbf{x}_0) &= \frac{P(\mathbf{x}_0|i)P(i)}{P(1)P(\mathbf{x}_0|1) + P(2)P(\mathbf{x}_0|2)} \\ &= \frac{f_i(\mathbf{x}_0)P(i)}{P(1)f_1(\mathbf{x}_0) + P(2)f_2(\mathbf{x}_0)} \end{aligned} \tag{2.12}$$

donde $i = 1, 2$

Se utilizará el concepto de función discriminante:

Definición

Se llama función discriminante, f_d aquella que:

$$f_d : \mathbf{\Omega} \longrightarrow \mathbb{R}, \tag{2.13}$$

donde $\mathbf{\Omega}$ es el espacio de observaciones posibles, definida de tal manera que si $f_d(\mathbf{x}_0) > 0 \Rightarrow \mathbf{x}_0 \in \pi_i$ y en caso contrario $\mathbf{x}_0 \notin \pi_i$.

Teniendo esta definición en mente para dos poblaciones y teniendo en cuenta el coste de clasificación errónea, se obtiene la siguiente función discriminante:

$$f_d(\mathbf{x}_0) = \frac{f_1(\mathbf{x}_0)P(1)}{c(1|2)} - \frac{f_2(\mathbf{x}_0)P(2)}{c(2|1)} \tag{2.14}$$

Si se sustituye f_1, f_2 por funciones normales se obtiene la función discriminante:

$$\begin{aligned} f_d(\mathbf{x}) = (\mathbf{x}-\mu_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1) + log\left(\frac{P(1)}{c(1|2)}\right) \\ -(\mathbf{x}-\mu_2)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2) - log\left(\frac{P(2)}{c(2|1)}\right) \end{aligned} \tag{2.15}$$

El análisis canónico de poblaciones busca identificar las direcciones de mayor diferenciación entre las distintas L poblaciones.

Definición

La covarianza intragrupo de un par de variables en la l-ésima población:

$$W_l(X_j, X_{j'}) = \frac{1}{N_l} \sum_{i \in I_l} (x_{ij} - \overline{x}_{jl})(x_{ij'} - \overline{x}_{j'l}) \tag{2.16}$$

Definición

La covarianza intergrupos

$$B(X_j, X_{j'}) = \sum_{l=1}^{L} \frac{N_l}{N} (\overline{x}_{jl} - \overline{x}_j)(\overline{x}_{j'l} - \overline{x}_{j'}) \tag{2.17}$$

Entonces se buscan combinaciones lineales que maximicen la varianza entre grupos respecto la varianza por tanto, se busca maximizar la función

$$f(\mathbf{a}) = \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{T}\mathbf{a}} \tag{2.18}$$

Tomando la restricción $\mathbf{a}^T\mathbf{T}\mathbf{a}$ y aplicando los multiplicadores de Lagrange, se obtiene que el vector $\mathbf{a}$ es el vector propio con valor propio máximo λ de la matriz $\mathbf{T}^{-1}\mathbf{B}$.

Contenidos

Métodos supervisados

Una red neuronal artificial es un modelo predictivo basado en el funcionamiento del cerebro humano.
Cada neurona recibe la información de las neuronas anteriores, las procesa y manda la señal de salida a las neuronas siguientes.
La ventaja es que este tipo de modelos se pueden escalar para tener la complejidad que requieran los datos a costa de perder interpretabilidad, es por ello que normalmente se utilizan con fines únicamente predictivos.

Métodos supervisados

Conceptos principales:

- Pesos sinápticos.
- Función de activación.
- Neurona artificial.

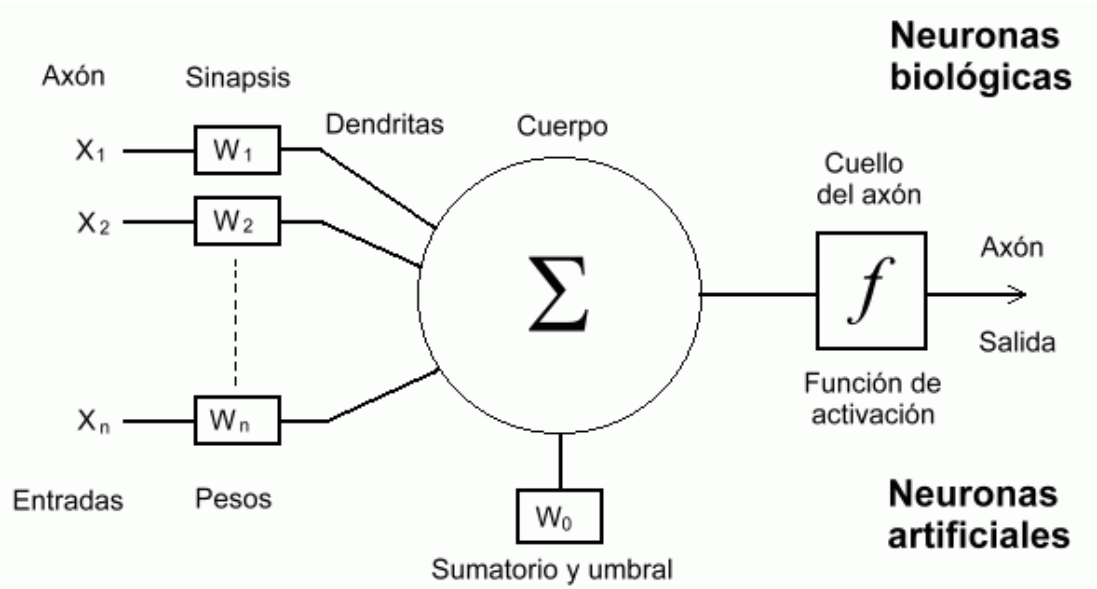


Figura: Analogía neurona biológica y artificial procedente de `https://www.um.es/LEQ/Atmosferas/Ch-VI-3/F63s4p3.htm`

Métodos supervisados

Tipos de capas:

- Capas de entrada.
- Capa oculta.
- Capa de salida.

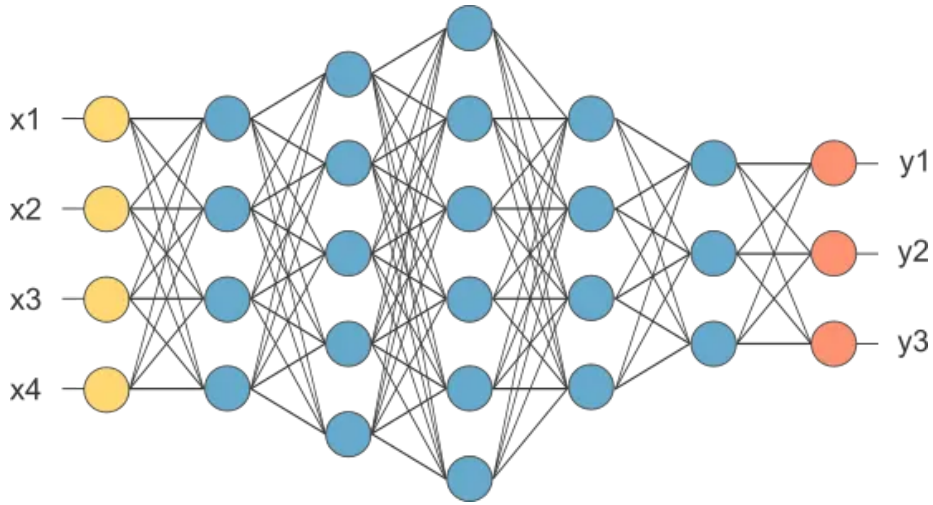


Figura: Estructura de una red neuronal como un grafo procedente de `https://www.neuraldesigner.com/learning/tutorials/neural-network`

Métodos supervisados

El ajuste de las redes neuronales se realiza con un proceso llamado *backpropagation*. El proceso consiste en los siguientes pasos

- Se inicia el modelo con un conjunto de parámetros que puede ser aleatorio o prefijados por el usuario.
- Se evalúa la red neuronal con esos parámetros.
- Se actualizan los parámetros de acuerdo a algún método de optimización como puede ser el método del gradiente.
- Se repiten los dos pasos anteriores hasta que se alcanza algún criterio de parada.

Contenidos

Métodos supervisados

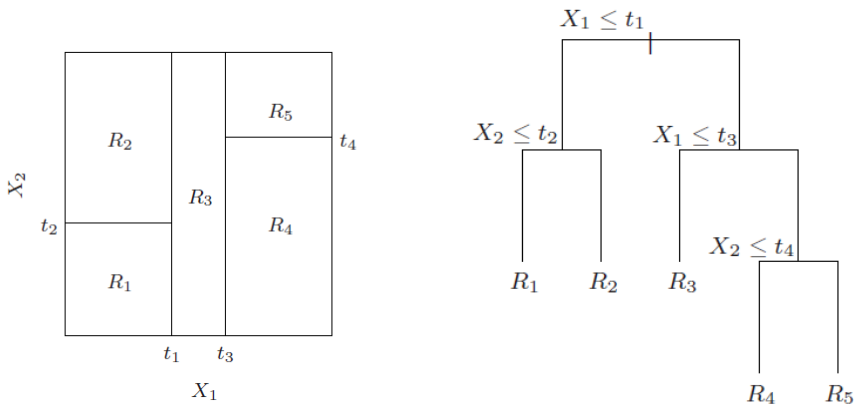
Definición

Los árboles de decisión son conjuntos de reglas discriminantes que fraccionan el espacio de observaciones en regiones donde se modeliza la variable respuesta de manera sencilla. En el proceso de ajuste generan un grafo de tipo árbol con un nodo raíz y los nodos hoja representan cada una de las regiones resultantes.

Métodos supervisados

Conceptos importantes:

- Partición de indices (j, s).
- Nodo terminal o nodo hoja.
- Nodo padre y nodo hijo.
- Profundidad y tamaño del árbol.



(a) División de $\mathbb{R}^p$ (b) Diagrama resultante

Figura: Representación de la división de $\mathbb{R}^p$ y el diagrama de árbol resultante.

Árboles de regresión

Sea el caso en el que la variable respuesta es cuantitiva y las variables predictoras son cuantitativas o cualitativas.
Tras haber ajustado el árbol hasta tener un tamaño $|T| = M$ resultando en las regiones $R_1, \ldots, R_M$. Entonces el predictor será el siguiente:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} \hat{c}_m \cdot \mathbf{1}_m(\mathbf{x}) \tag{2.19}$$

donde $\mathbf{1}_m(\mathbf{x})$ es la función característica de la región R_m, $\forall m = 1, \ldots, M$. Y la constante $\hat{c}_m$ es la media de la variable respuesta en la región R_m.

Árboles de regresión

Una partición (j, s) que particione el espacio en las dos regiones R_{m_1}, R_{m_2} es elegida si minimiza la siguiente cantidad:

$$\sum_{i/\mathbf{x}_i \in R_{m_1}} (y_i - \hat{c}_{m_1})^2 + \sum_{i/\mathbf{x}_i \in R_{m_2}} (y_i - \hat{c}_{m_2})^2 \tag{2.20}$$

Árboles de regresión

Se puede definir la medida de bondad de ajuste a los datos de un determinado árbol de la siguiente manera:

$$Q(T) = \sum_{m=1}^{M} \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2, \tag{2.21}$$

Si se hace demasiado grande el árbol, se puede dar una situación de sobreajuste, en la que el error medio cuando se ajuste sea pequeño pero que el modelo pierda capacidad predictiva.

Árboles de regresión

Definición

Se llama poda al proceso en el que dado un árbol inicial de tamaño T_0 se revierten ciertas particiones terminales que no aportan en la relación coste-complejidad, es decir, aumentan demasiado la complejidad (aumentando la varianza), sin reducir el coste en exceso.

Se puede añadir un término regularizador para el error cuadrático.

$$Q_\alpha(T) = \sum_{m=1}^{M} \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2 + \alpha|T|, \qquad (2.22)$$

Árboles de clasificación

Sea las variables predictoras como antes, pero la variable respuesta ahora es cualitativa. Supóngase que puede tomar L valores distintos que llamaremos l, $l = 1, \ldots, L$.

$\hat{p}_{lm}$ = Proporción de observaciones en las que $y_i = l$, en la región R_m. (2.23)

Árboles de clasificación

Teniendo en cuenta esto se define lo siguiente:

- Impureza de un nodo; $1 - \text{máx}_{l \in L}\, \hat{p}_{lm}$
- Índice Gini $G = \sum_{l=1}^{L} \hat{p}_{lm}(1 - \hat{p}_{lm})$
- Entropía de un nodo
  $H(\hat{p}_{lm}) = -\hat{p}_{lm} log(\hat{p}_{lm}) - (1 - \hat{p}_{lm}) log(1 - \hat{p}_{lm})$
- Ganancia $Ganancia = H(padre) - \sum_{i=1}^{k} \frac{N_i}{N_{padre}} H_i, 1$
- Ratio de ganancia $\frac{Ganancia}{H_{padre}}$

Árboles

Los distintos métodos de elección de las particiones dan lugar a los distintos algoritmos:

- *CART* de Breiman para árboles tanto de regresión como de clasificación.
- *C.4.5* de Quinlan también para cualquier modalidad
- *CHAID* de Kass únicamente para árboles de clasificación.

Sesgo y varianza de un modelo

Definición

Se llama error de predicción esperado de la observación $\mathbf{x}_0$ a la siguiente expresión:

$$EPE(\mathbf{x}_0) = \mathbb{E}((Y - \hat{f}(\mathbf{x}_0))^2) \tag{2.24}$$

Proposición

El error de predicción esperado se puede dividir en un termino irreducible, el sesgo del modelo y la varianza:

$$EPE(\mathbf{x}_0) = Sesgo(\hat{f}(\mathbf{x}_0))^2 + Var(\hat{f}(\mathbf{x}_0)) \tag{2.25}$$

donde el $Sesgo(\hat{f}(\mathbf{x}_0)) = \mathbb{E}(\hat{f}(\mathbf{x}_0)) - f(\mathbf{x}_0)$.

Sesgo y varianza de un modelo

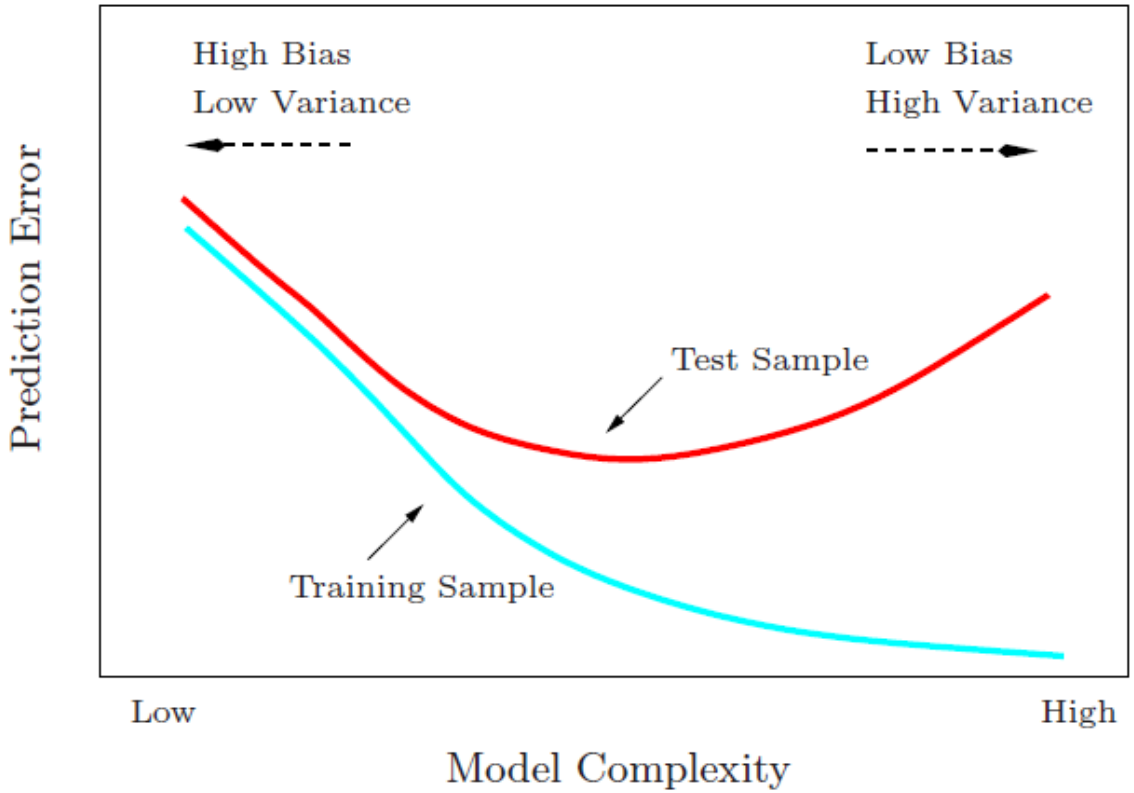


Figura: Balance sesgo varianza

Bosques Aleatorios

Los bosques aleatorios es un técnica que definió Breiman. Este proceso busca hacer crecer varios árboles de manera totalmente aleatoria concretamente:

- En cada árbol se utiliza una muestra aleatoria sin reemplazamiento del conjunto de datos iniciales.
- Se establece un número j_{try} por el usuario. Para cada partición a realizar se escogen de manera aleatoria j_{try} variables y se elige la que más se ajusta al criterio dado. Y se hace crecer el árbol hasta que se da un criterio de parada.
- Una vez crecidos los árboles, se toma como predicción de una nueva observación la media de las predicciones de los datos en el caso de que la variable respuesta sea continua. En el caso contrario, se toma el voto por mayoría.

Contenidos

Métodos no supervisados

Los métodos no supervisados son aquellos que buscan analizar la estructura subyacente de los datos. De esta manera no hay una diferenciación de variables, sino que se analizan las relaciones entre las variables.

Contenidos

Dado un vector aleatorio $\mathbf{x}$ de longitud p. Sea $r = rg(\mathbf{X})$. Entonces se definen las componentes principales:

Definición

Se definen las componentes principales

$$Z_j = v_{1j}X_1 + \ldots v_{pj}X_p = \mathbf{v}_j^T \mathbf{x}^T \quad j = 1 \ldots r \tag{3.1}$$

Donde $\mathbf{v}_j$ es un vector columna con p escalares y la nueva variable aleatoria Z_j cumple lo siguiente:

- Si $j = 1$ $Var(Z_1)$ es máxima restringido a $\mathbf{v}_1^T \mathbf{v}_1 = 1$
- Si $j > 1$ debe cumplir:
  - $Cov(Z_j, Z_i) = 0 \quad \forall i \neq j$
  - $\mathbf{v}_j^T \mathbf{v}_j = 1$
  - $Var(Z_j)$ es máxima.

Definición

Dada una matriz $\mathbf{X} \in \mathbb{M}_{N \times p}(\mathbb{R})$ existe la descomposición en valores singulares :

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T \tag{3.2}$$

Donde:

- $\mathbf{U}$ matriz ortogonal y de tamaño $N \times r$
- $\mathbf{D}$ matriz de tamaño $r \times r$ diagonal, cuyos elementos no nulos son los valores singulares $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$, es decir

$$\mathbf{D} = \begin{pmatrix} \sigma_1 & 0 & \cdots & \cdots \\ 0 & \sigma_2 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & 0 & \sigma_r \end{pmatrix} \tag{3.3}$$

- $\mathbf{V}$ matriz ortogonal y de tamaño $p \times r$.

Métodos no supervisados

Las componentes principales se pueden calcular de dos maneras:

- Descomposición en vectores y valores propios de la matriz de covarianzas.
- Descomposición en valores singulares de la matriz de datos

Definición

Sea $\mathbf{A} \in \mathbb{M}_{N\times p}(\mathbb{R})$. Definimos la norma de Frobenius de la matriz $\mathbf{A}$ como:

$$||\mathbf{A}||_F = (tr(\mathbf{A}^T\mathbf{A}))^{\frac{1}{2}} = \left(\sum_{i=1}^{N}\sum_{j=1}^{p} a_{ij}^2\right)^{\frac{1}{2}} \tag{3.4}$$

Proposición

La norma de Frobenius es invariante a transformaciones ortogonales.

Definición

Se llama matriz reducida de orden $m \leq r$ de $\mathbf{X}$ y se denota como $\mathbf{X}_m$, a la matriz $N \times m$ resultado de:

$$\mathbf{X}_m = \mathbf{U}_m \mathbf{D}_m \mathbf{V}_m^T \tag{3.5}$$

Donde:

- $\mathbf{U}_m$ matriz ortogonal de tamaño $N \times m$, resultado de tomar de $\mathbf{U}$ únicamente la matriz las m primeras columnas.
- $\mathbf{D}_m$ matriz cuadrada de tamaño m diagonal con los m primeros valores singulares.
- $\mathbf{V}_m$ matriz ortogonal de tamaño $p \times m$ obtenida al tomar las m primeras columnas de $\mathbf{V}$.

Teorema (De Eckart-Young)

Sea $\mathbf{A}$ una matriz de coeficientes reales de tamaño $N \times p$ y rango r. Entonces se cumple que :

$$||\mathbf{A} - \mathbf{B}||_F \leq ||\mathbf{A} - \mathbf{A}_m||_F \quad \forall \mathbf{B}/rg(\mathbf{B}) = m \leq r \tag{3.6}$$

El número m de componentes principales se puede elegir teniendo en cuenta la siguiente cantidad:

$$t_m = \frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{r} \lambda_j} \tag{3.7}$$

La proporción de variabilidad acumulada por las m primeras componentes.

Contenidos

Métodos no supervisados

El objetivo principal del análisis factorial es estudiar la existencia de estructuras latentes en conjuntos de datos
Esto se traduce buscando si las variables aleatorias pueden ver su variabilidad explicada por un conjunto menor de factores comunes.
Por tanto, lo que se busca es expresar la variabilidad del conjunto de datos en términos comunes con el resto y una parte específica.

Métodos no supervisados

Supóngase un vector aleatorio $\mathbf{x} = [X_1, \ldots, X_p]$ de longitud p con una distribución $N(0, \Sigma)$, centrada sin pérdida de generalidad, donde la matriz de covarianzas $\mathbf{\Sigma}$ es simétrica y definido positiva. El modelo factorial se establece:

$$X_j = \lambda_{1j} F_1 + \ldots + \lambda_{mj} F_m + \psi_j U_j \quad j = 1 \ldots p \tag{3.8}$$

Donde:

- λ_{kj} es la saturación de la j-ésima variable en el factor común k-ésimo.
- F_k es el k-ésimo factor común
- ψ_j es la saturación específica de la variable X_j en el factor único.
- U_j es el factor específico para la variable X_j.

Métodos no supervisados

El análisis factorial se apoya en las siguientes hipótesis :

- Los factores comunes F_k son variables aleatorias que siguen una distribución marginal $N(0,1)$. Además se supondrá que $Cov(F_k, F_{k'}) = 0, k \neq k', \forall k, k' = 1, \ldots, m$. Se suponen completamente independientes de los factores específicos.
- Los factores específicos U_j son variables aleatorias con una distribución normal $N(0,1)$ no correladas. Se suponen completamente independientes de los factores comunes.

Definición

Llamaremos matriz de saturaciones de los factores o matriz de saturaciones, $\mathbf{\Lambda}$ de tamaño $p \times m$ a la matriz :

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{11} & \cdots & \lambda_{1m} \\ \vdots & \ddots & \vdots \\ \lambda_{p1} & \cdots & \lambda_{pm} \end{pmatrix} \tag{3.9}$$

Definición

Se llama comunalidad de la variable X_j, a $h_j^2 = \sum_{k=1}^{m} \lambda_{kj}^2$.

Definición

Llamaremos matriz específica a la matriz diagonal $\mathbf{\Psi}$ de tamaño $p \times p$ a aquella que contiene los términos ψ_j donde $j = 1, \ldots, p$:

$$\mathbf{\Psi} = \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{pmatrix} \tag{3.10}$$

Definición

Se llama especificidad o unicidad de la variable X_j a $\psi_j^2, \forall j = 1, \ldots, p$.

Proposición

La varianza de la variable aleatoria X_j, σ_j^2, se puede descomponer de la siguiente manera:

$$\sigma_j^2 = \sum_{k=1}^{m} \lambda_{kj}^2 + \psi_j^2 \quad \forall j = 1, \ldots, p \tag{3.11}$$

Proposición

La covarianza de dos variables $Cov(X_j, X_{j'})$, $\sigma_{jj'}$ cumple que

$$\sigma_{jj'} = \sigma_{j'j} = \sum_{k=1}^{m} \lambda_{kj}\lambda_{kj'} \tag{3.12}$$

Teorema

La matriz $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}^2$

Para la estimación de la matriz de satuaciones hay que tener en cuenta la siguiente descomposición:

$$\mathbf{S} = \mathbf{\Lambda}\mathbf{\Lambda}^T + \mathbf{\Psi}^2 \tag{3.13}$$

Por tanto, la matriz de saturaciones $\mathbf{\Lambda}$ se puede estimar haciendo la descomposición en valores y vectores propios de esta matriz.

$$\mathbf{S} - \mathbf{\Psi}^2 = \mathbf{\Lambda}\mathbf{\Lambda}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{3.14}$$

Si $\Psi = 0$ se tiene la descomposición en valores y vectores propios de la matriz de covarianzas muestral.

Métodos no supervisados

Para poder calcular simultáneamente la matriz de saturaciones la matriz específica se plantea el siguiente proceso iterativo. El método se inicia con $\hat{\mathbf{\Psi}}_0 = 0$ de tal manera que $\mathbf{S}_0^* = \mathbf{S}$. En cada iteración se efectúan las siguientes operaciones:

1. Se calcula $\mathbf{S}_i^* = \mathbf{S} - \hat{\mathbf{\Psi}}_i^2$.
2. Se calcula la descomposición $\hat{\mathbf{\Lambda}}_i$ como se ha detallado antes obteniendo $\mathbf{U}_i$ y $\mathbf{D}_i$ de tamaños $p \times m$ y $m \times m$.
3. Se obtiene la nueva estimación de la $\hat{\mathbf{\Psi}}_{i+1}^2 = \mathbf{S} - \hat{\mathbf{\Lambda}}_i \hat{\mathbf{\Lambda}}_i^T$

El modelo está indeterminado por transformaciones ortogonales.

$$\mathbf{x}^T = (\mathbf{\Lambda}\mathbf{T}^T)(\mathbf{T}\mathbf{f}^T) + \Psi\mathbf{u}^T \tag{3.15}$$

Para evitar estas indeterminaciones se suelen imponer dos restricciones:

$\mathbf{\Lambda}^T\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ ó $\mathbf{\Lambda}^T\mathbf{\Lambda}$ son diagonales con los valores en orden decreciente (3.16)

Métodos no supervisados

Para interpretar el modelo factorial hay que tener en cuenta los siguientes aspectos:

- Proporción de varianza común explicada.
- Especificidad de cada una de las variables.
- Saturaciones de cada variable en los factores.
- Significado de los factores

Una vez interpretados los resultados, se puede llegar a los objetivos del análisis factorial, estudiar la estructura latente de los datos en términos de variabilidades comunes.

Contenidos

Métodos no supervisados

Objetivo: Agrupar en grupos homogéneos o *clusters* las observaciones o mediciones.

Definición

Diremos que un *cluster* o conglomerado es un subconjunto de las observaciones que son similares entre sí.

Se dice que dos observaciones $\mathbf{x}_i, \mathbf{x}_{i'}$ están relacionadas, $\mathcal{R}$, cuando pertenecen al mismo *cluster*.

Definición

Se llama *clustering* a la partición que provoca la relación $\mathcal{R}$ del espacio de observaciones.

Métodos no supervisados

Tipos de algoritmos de *clustering*:

- **Particional** una única partición que se va modificando
- **Jerárquico** otorga las posibles particiones desde *N cluster* hasta un sólo *cluster*.
  - **Aglomerativo:** se empiezan con *N clusters* con una observación en cada uno y se van juntando según un criterio establecido.
  - **Divisivo** se empiezan con un *cluster* con todas las observaciones y se va dividiendo según algún criterio establecido.

Dependiendo de cómo se definan las distancias entre los *clusters* se obtendrá un algoritmo u otro:

- El vecino más proximo: $d(C, C') = \text{mín}_{\mathbf{x}_i \in C, \mathbf{x}_{i'} \in C'}(\mathbf{x}_i, \mathbf{x}_{i'})$.
- El vecino más lejano: $d(C, C') = \text{máx}_{\mathbf{x}_i \in C, \mathbf{x}_{i'} \in C'}(\mathbf{x}_i, \mathbf{x}_{i'})$.
- Enlace medio

$$d(C, C') = \frac{1}{N_C} \frac{1}{N'_C} \sum_{\mathbf{x}_i \in C} \sum_{\mathbf{x}_{i'} \in C'} d(\mathbf{x}_i, \mathbf{x}_{i'}). \tag{3.17}$$

- Ward:

$$d(C, C') = \frac{1}{N_C} \frac{1}{N'_C} \sum_{\mathbf{x}_i \in C} \sum_{\mathbf{x}_{i'} \in C'} (d(\mathbf{x}_i, \mathbf{x}_{i'}))^2 \tag{3.18}$$

Métodos no supervisados

El proceso que siguen los algoritmos jerárquicos aglomerativos es el siguiente:

- Se inicia con una partición que tiene *N clusters* con una única observación cada uno.
- Se calcula la matriz de distancias entre los *clusters*.
- Se unen los dos clusters que menor distancia tengan entre ellos. De esta manera, se reduce el número de clusters en 1.
- Se repiten los dos pasos anteriores hasta tener un único cluster.

Métodos no supervisados

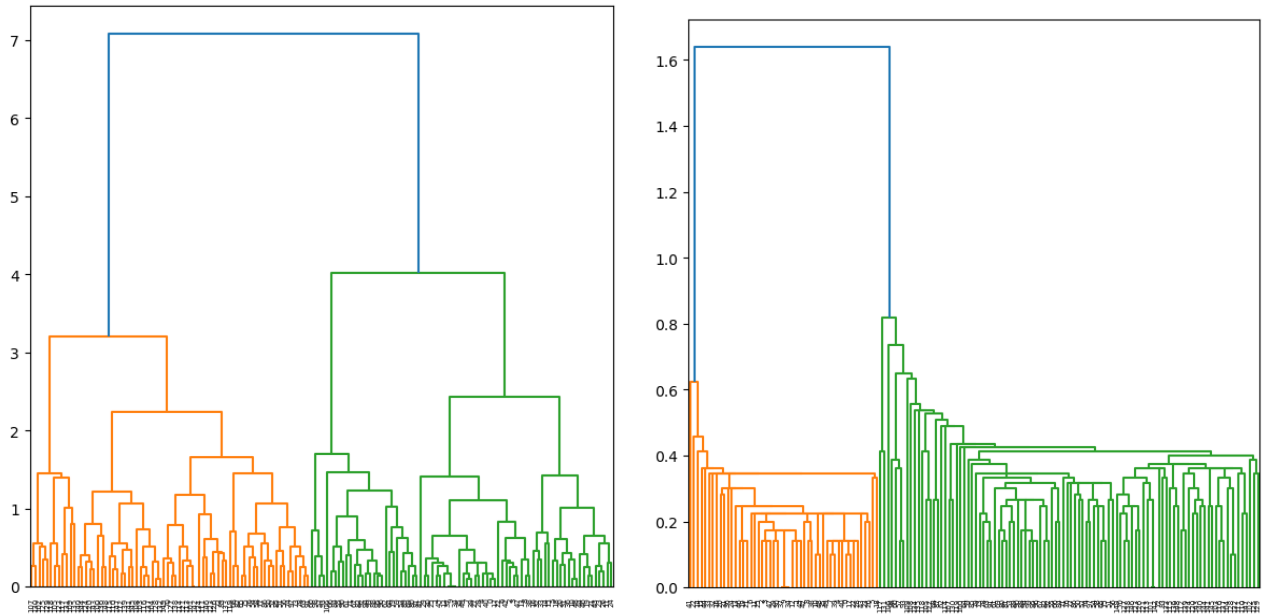
Estos algoritmos tienen la característica que permiten representar sus resultados como un dendograma.

Definición

Un dendograma es un diagrama de árbol en el que en el eje horizontal se sitúan las observaciones mientras que en el eje vertical se representan las distancias. Cada nodo representa una unión de los clusters.

Métodos no supervisados

Los siguientes dendogramas 5a 5b son resultado de aplicar un algoritmo jerárquico al conjunto *IRIS* de Fisher.



(a) Enlace completo (b) Enlace simple

Figura: Diagramas obtenidos utilizando la biblioteca de Python Scikit-Learn

Como se ha dicho, los algoritmos particionales generan un único *clustering* que se va modificando a lo largo de la ejecución del algoritmo.
Para poder dar el algoritmo más importante de este tipo se necesita del siguiente concepto:

Definición

Se llama *centroide* de un cluster C_k, $k = 1, \ldots, K$ al vector de tamaño p cuyas componentes son las medias de cada una de las variables de todas las observaciones del *cluster* C_k:

$$\overline{x}_{jk} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij} \quad \forall j = 1, \ldots, p \tag{3.19}$$

Métodos no supervisados

El algoritmo de K-medias se desarrolla de la siguiente manera:

1. Se asignan a cada una de las observaciones en cada uno de los K *clusters*. Se calculan los centroides de dichos *clusters*.
2. Se comprueba la distancia euclídea de cada una de las observaciones a los centroides de los *clusters* y se reasignan las observaciones de acuerdo al más cercano. Tras esto se recalculan los centroides.
3. Se repiten los dos pasos anteriores hasta que las asignaciones no cambien.

Contenidos

Aplicación

Nombre	Tipo de dato	Medida	Descripción
Cemento (variable 1)	continuo	kg en una mezcla m^3	Cantidad de cemento presente en la mezcla
Escoria de alto horno (variable 2)	continuo	kg en una mezcla m^3	Producto resultante de distintas fundiciones Mejora las capacidades hidráulicas.
Ceniza volante (variable 3)	continuo	kg en una mezcla m^3	Sustitutivo cemento.
Agua (variable 4)	continuo	kg en una mezcla m^3	Agua en la mezcla.
Superplastificante (variable 5)	continuo	kg en una mezcla m^3	Mejora la resistencia
Agregado grueso (variable 6)	continuo	kg en una mezcla m^3	Abaratador de la mezcla
Agregado fino (variable 7)	continuo	kg en una mezcla m^3	Abaratador de la mezcla
Edad	continuo	Día (1-365)	Tiempo desde mezcla.
Resistencia a la compresión	continuo	MPa	

Cuadro: Resumen de las variables estudiadas

El tamaño de la muestra es de 1030 observaciones.

Aplicación

Objetivos:

- Detallar y analizar patrones de comportamiento en los cementos en base a las características descritas.
- Crear y analizar distintos modelos que permitan la predicción de la resistencia a la compresión.

Aplicación

	Cemento (kg en una mezcla m3)	Escoria de alto horno (kg en una mezcla m3)	Ceniza volante (kg en una mezcla m3)	Agua (kg en una mezcla m3)	Superplastificante(kg en una mezcla m3)	Agregado grueso(kg en una mezcla m3)	Agregado fino (kg en una mezcla m3)	Edad (1~365 días)	Resistencia a la compresión del concreto (MPa)
count	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000
mean	281.167864	73.895825	54.188350	181.567282	6.204660	972.918932	773.580485	45.662136	35.817961
std	104.506364	86.279342	63.997004	21.354219	5.973841	77.753954	80.175980	63.169912	16.705742
min	102.000000	0.000000	0.000000	121.800000	0.000000	801.000000	594.000000	1.000000	2.330000
25%	192.375000	0.000000	0.000000	164.900000	0.000000	932.000000	730.950000	7.000000	23.710000
50%	272.900000	22.000000	0.000000	185.000000	6.400000	968.000000	779.500000	28.000000	34.445000
75%	350.000000	142.950000	118.300000	192.000000	10.200000	1029.400000	824.000000	56.000000	46.135000
max	540.000000	359.400000	200.100000	247.000000	32.200000	1145.000000	992.600000	365.000000	82.600000

Figura: Análisis descriptivo de los datos

Analisis de componentes principales:

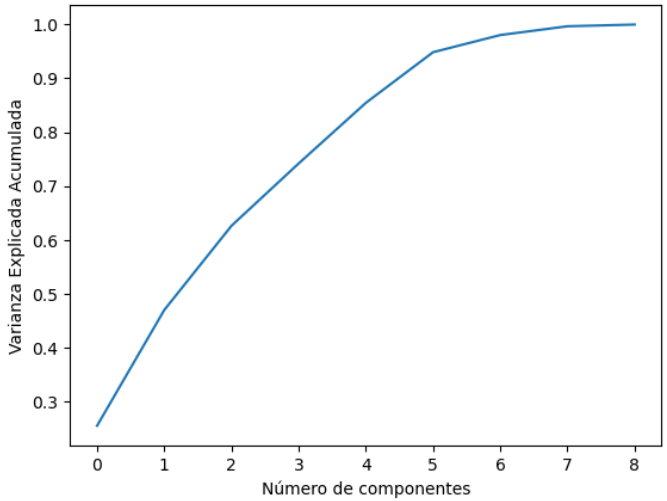


Figura: Proporción de varianza acumulada explicada por las componentes principales.

En particular la varianza explicada por cada una de las componentes:

$$(0{,}25 \;\; 0{,}22 \;\; 0{,}16 \;\; 0{,}12 \;\; 0{,}11 \;\; 0{,}09 \;\; 0{,}03 \;\; 0{,}02 \;\; 0.) \qquad (4.1)$$

Aplicación

Las componentes principales tienen las cargas:

$$Z_1 = (0,04; 0,16; -0,37; 0,56; -0,54; 0,06; -0,38; 0,26; -0,11)\mathbf{x}^T \tag{4.2}$$

$$Z_2 = (0,54; 0,14; -0,27; -0,12; 0,25; -0,22; -0,19; 0,25; 0,63)\mathbf{x}^T \tag{4.3}$$

$$Z_3 = (0,36; -0,7; 0,02; -0,12; -0,19; 0,55; 0; 0,17; 0,03)\mathbf{x}^T \tag{4.4}$$

Aplicación

Modelo lineal de regresión:

$$Y = -51{,}45 + 0{,}13X_1 + 0{,}11X_2 + 0{,}1X_3 - 0{,}12X_4 + 0{,}26X_5 + 0{,}03X_6 + 0{,}03X_7 + 0{,}11X_8 \tag{4.5}$$

Resultados:

- Conjunto de entrenamiento
  - **MSE:** $108,35$
  - R^2 : $0{,}61$
- Conjunto de prueba
  - **MSE:** $106,02$
  - R^2 : $0{,}61$

Aplicación

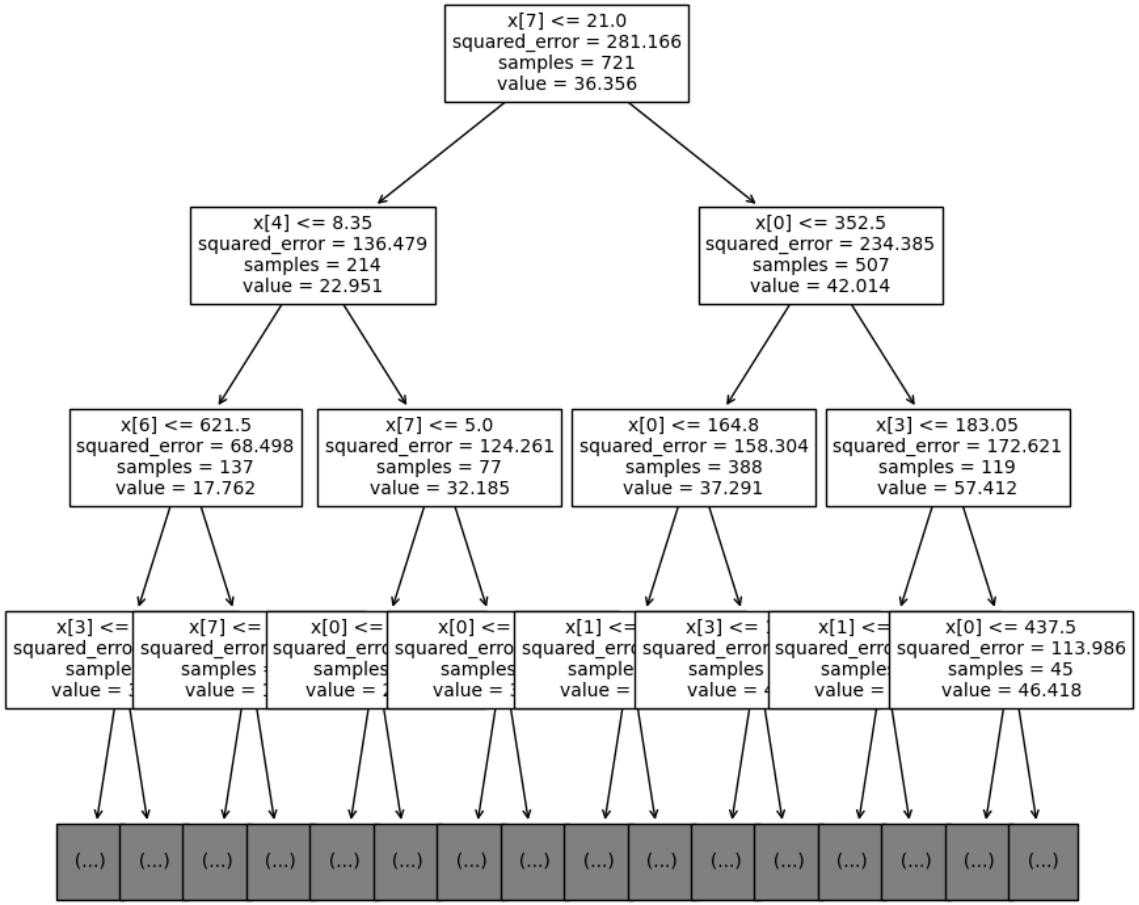


Figura: Diagrama resultante hasta profundidad 3

Aplicación

Árbol de regresión:

- Conjunto de entrenamiento
  - **MSE:** $1,15$
  - R^2 : $0,996$
- Conjunto de prueba
  - **MSE:** $53,28$
  - R^2 : $0,803$

Observación: las medidas del conjunto de prueba nos dice que estamos ante un caso en el que se ha producido un sobreajuste.

Aplicación

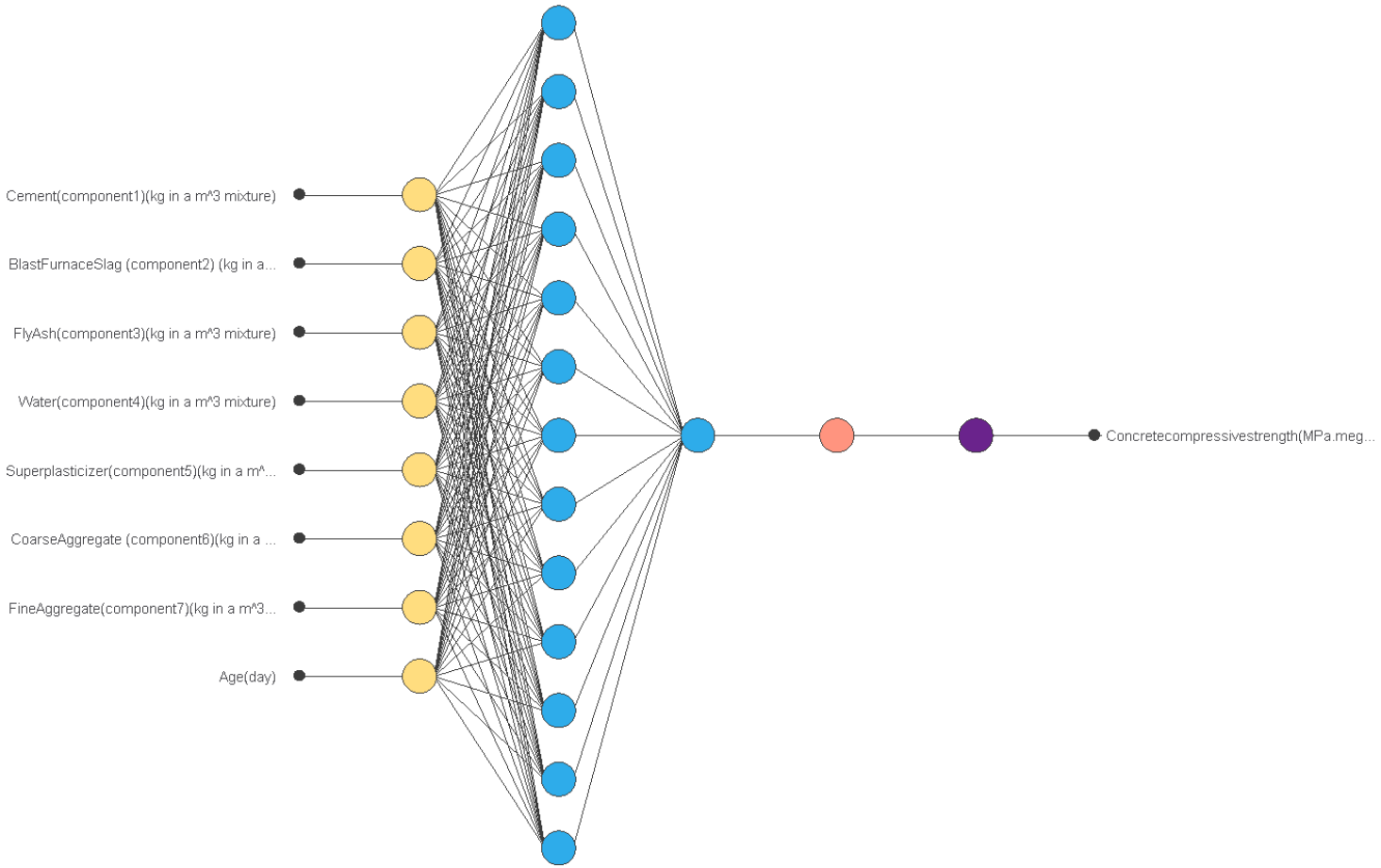


Figura: Estructura de la red neuronal, realizada con el programa Neural Designer.

Aplicación

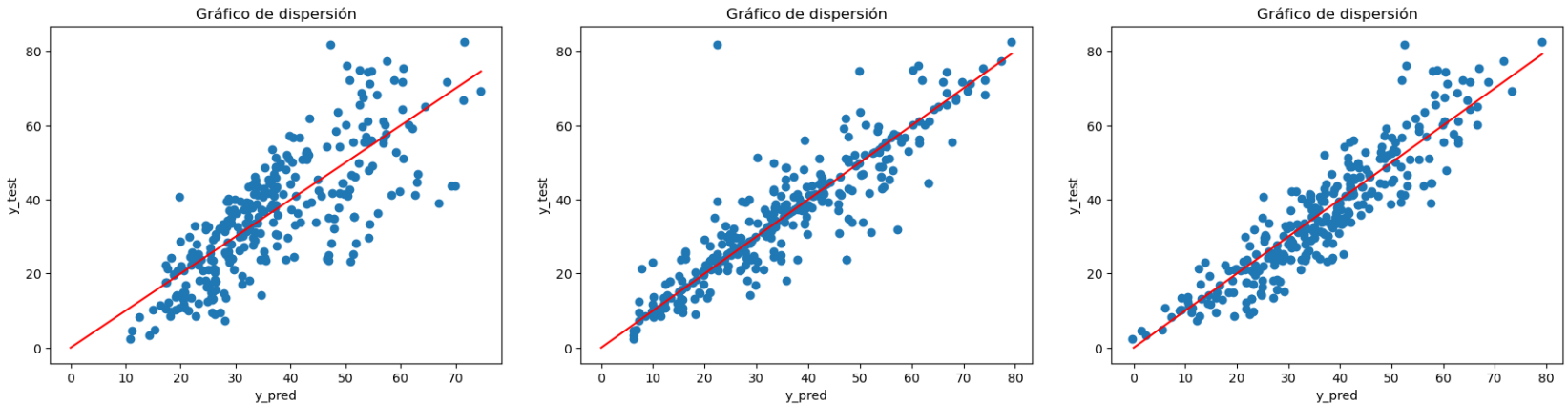
Red neuronal:

- Conjunto de entrenamiento
  - **MSE:** $45,748$
  - R^2 : $0,837$
- Conjunto de prueba
  - **MSE:** $51,805$
  - R^2 : $0,809$

Observación: los resultados obtenidos con el software Neural Designer son mejores.

Aplicación

Comparación modelos:



(a) Regresión Lineal (b) Árbol de regresión (c) Red neuronal

Figura: Gráficos de comparación de los valores predichos con el valor real.