

Revisión de métodos multivariantes supervisados y no supervisados

Pedro Ángel Fraile Manzano

6 de julio de 2023

1 Introducción

2 Métodos supervisados

- Métodos lineales para regresión
 - Métodos de ajuste
 - Inferencias estadísticas sobre $\hat{\beta}$
 - Regresión multivariante
 - Selección de subconjuntos y métodos penalizados
- Clasificación
 - Funciones discriminantes
 - Análisis canónico de poblaciones
- Redes neuronales
- Árboles de decisión y bosques aleatorios.
 - Árboles de regresión

3 Métodos no supervisados

1 Introducción

2 Métodos supervisados

- Métodos lineales para regresión
 - Métodos de ajuste
 - Inferencias estadísticas sobre $\hat{\beta}$
 - Regresión multivariante
 - Selección de subconjuntos y métodos penalizados
- Clasificación
 - Funciones discriminantes
 - Análisis canónico de poblaciones
- Redes neuronales
- Árboles de decisión y bosques aleatorios.
 - Árboles de regresión

3 Métodos no supervisados

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.

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

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.

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.

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.

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.

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.

Objetivo:

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

1 Introducción

2 Métodos supervisados

- Métodos lineales para regresión
 - Métodos de ajuste
 - Inferencias estadísticas sobre $\hat{\beta}$
 - Regresión multivariante
 - Selección de subconjuntos y métodos penalizados
- Clasificación
 - Funciones discriminantes
 - Análisis canónico de poblaciones
- Redes neuronales
- Árboles de decisión y bosques aleatorios.
 - Árboles de regresión

3 Métodos no 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:

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 $\mathbf{x} = [X_1, \dots, X_p]$

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

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 , $i = 1, \dots, N$ 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 , $i = 1, \dots, N$ representan una observación de las K variables respuesta.

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 $\hat{\mathbf{y}}_0$.

1 Introducción

2 Métodos supervisados

• Métodos lineales para regresión

- Métodos de ajuste
- Inferencias estadísticas sobre $\hat{\beta}$
- Regresión multivariante
- Selección de subconjuntos y métodos penalizados

• Clasificación

- Funciones discriminantes
- Análisis canónico de poblaciones

• Redes neuronales

- Árboles de decisión y bosques aleatorios.
 - Árboles de regresión

3 Métodos no 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.

Tomamos el modelo en el que:

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

donde ε cumple:

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

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

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

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

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

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

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

$$\hat{\mathbf{y}} = \mathbf{X}\beta \quad (2.4)$$

Definición

Llamamos error de predicción cuadrático entre una variable observable Y y la predicción obtenida por un cierto modelo \hat{Y} a la expresión $(Y - \hat{Y})^2$

Para medir la bondad de ajuste al tomar una muestra de N observaciones se puede utilizar la suma de errores cuadráticos, RSS

$$RSS(\beta) = \sum_{i=1}^N (y_i - \mathbf{x}_i\beta)^2 = (\mathbf{y} - \mathbf{X}\beta)(\mathbf{y} - \mathbf{X}\beta)^T \quad (2.5)$$

Se puede obtener una estimación de β minimizando RSS respecto de β , obteniéndose que en el caso de que \mathbf{X} sea de rango máximo la siguiente expresión:

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

Proposición

Con los supuestos del modelo, la estimación mediante mínimos cuadrados es equivalente al método de máxima verosimilitud

Por tanto, el vector de predicciones se puede expresar de la siguiente manera:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (2.7)$$

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 $\hat{\sigma}^2 \sim \frac{\sigma^2}{N - p - 1} \chi_{N-p-1}^2$

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

$$F = \frac{\frac{(RSS_0 - RSS_1)}{p_1 - p_0}}{\frac{RSS_1}{N - p_1 - 1}} \sim F_{(p_1 - p_0), (N - p_1 - 1)} \quad (2.8)$$

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

$$Y_k = \beta_{0k} + \sum_{j=1}^p X_j \beta_{jk} + \varepsilon_k = f_k(\mathbf{x}) + \varepsilon_k, \quad k = 1, \dots, K \quad (2.9)$$

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

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \quad (2.10)$$

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:

$$RSS(\mathbf{B}, \mathbf{\Sigma}) = \sum_{i=1}^N (\mathbf{y}_i - f(\mathbf{x}_i)) \mathbf{\Sigma}^{-1} (\mathbf{y}_i - f(\mathbf{x}_i))^T \quad (2.11)$$

En la anterior expresión aparece la distancia de Mahalanobis que se define de la siguiente manera:

Definición

La *distancia de Mahalanobis* entre dos observaciones $\mathbf{x}_i, \mathbf{x}_j$ extraídas de una misma población con matriz de covarianzas $\mathbf{\Sigma}$ se define de la siguiente manera:

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)^T} \quad (2.12)$$

Observación: Se puede definir de manera análoga en el caso de que se conozca la matriz de covarianzas muestral \mathbf{S} .

Se puede utilizar el estadígrafo F de la ecuación 2.8 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 $\lambda > 0$:

$$PRSS(\beta) = \sum_{i=1}^n (\mathbf{y}_i - \mathbf{x}_i \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (2.13)$$

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

1 Introducción

2 Métodos supervisados

- Métodos lineales para regresión
 - Métodos de ajuste
 - Inferencias estadísticas sobre $\hat{\beta}$
 - Regresión multivariante
 - Selección de subconjuntos y métodos penalizados
- Clasificación
 - Funciones discriminantes
 - Análisis canónico de poblaciones
- Redes neuronales
- Árboles de decisión y bosques aleatorios.
 - Árboles de regresión

3 Métodos no supervisados

La clasificación es el caso donde la variable respuesta es categórica o cualitativa.

La clasificación se utiliza

La clasificación se utiliza

1 Introducción

2 Métodos supervisados

- Métodos lineales para regresión
 - Métodos de ajuste
 - Inferencias estadísticas sobre $\hat{\beta}$
 - Regresión multivariante
 - Selección de subconjuntos y métodos penalizados
- Clasificación
 - Funciones discriminantes
 - Análisis canónico de poblaciones
- **Redes neuronales**
- Árboles de decisión y bosques aleatorios.
 - Árboles de regresión

3 Métodos no 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.

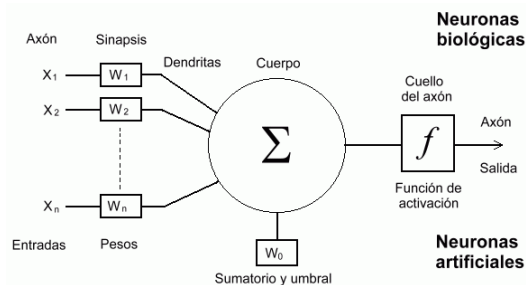


Figura: Analogía neurona biológica y artificial procedente de <https://www.um.es/LEQ/Atmosferas/Ch-VI-3/F63s4p3.htm>

Métodos supervisados

Tipos de capas:

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

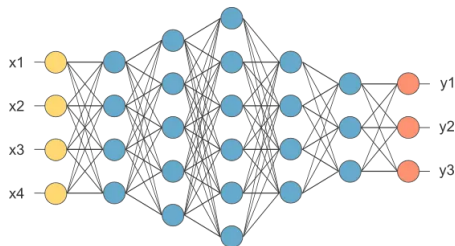


Figura: Estructura de una red neuronal como un grafo procedente de <https://www.neuraldesigner.com/learning/tutorials/neural-network>

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.

1 Introducción

2 Métodos supervisados

- Métodos lineales para regresión
 - Métodos de ajuste
 - Inferencias estadísticas sobre $\hat{\beta}$
 - Regresión multivariante
 - Selección de subconjuntos y métodos penalizados
- Clasificación
 - Funciones discriminantes
 - Análisis canónico de poblaciones
- Redes neuronales
- Árboles de decisión y bosques aleatorios.
 - Árboles de regresión

3 Métodos no supervisados

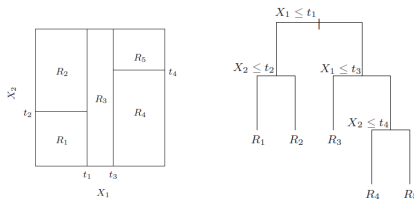
Definición

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 índices (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 cuantitativa 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, \dots, R_M . Entonces el predictor será el siguiente:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^M \hat{f}_m(\mathbf{x}) \cdot \mathbf{1}_m(\mathbf{x}) \quad (2.14)$$

donde $\mathbf{1}_m(\mathbf{x})$ es la función característica de la región R_m , $\forall m = 1, \dots, M$.

Árboles de regresión

Si en cada región queremos que \hat{f}_m sea una constante, \hat{c}_m . Utilizando el método de los mínimos cuadrados teniendo en cuenta la restricción de que sea constante se obtiene que:

$$\hat{c}_m = \frac{1}{N_m} \sum_{i/\mathbf{x}_i \in R_m} y_i \quad (2.15)$$

donde N_m es la cantidad de observaciones que hay en la región $R_m \ \forall m = 1, \dots, M$

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:

$$\sum_{i/\mathbf{x}_i \in R_{m_1}} (y_i - \hat{c}_{m_1})^2 + \sum_{i/\mathbf{x}_i \in R_{m_2}} (y_i - \hat{c}_{m_2})^2 \quad (2.16)$$

1 Introducción

2 Métodos supervisados

- Métodos lineales para regresión
 - Métodos de ajuste
 - Inferencias estadísticas sobre $\hat{\beta}$
 - Regresión multivariante
 - Selección de subconjuntos y métodos penalizados
- Clasificación
 - Funciones discriminantes
 - Análisis canónico de poblaciones
- Redes neuronales
- Árboles de decisión y bosques aleatorios.
 - Árboles de regresión

3 Métodos no supervisados

Métodos no supervisados