

Revisión De Técnicas De Clasificación Supervisada Y Su Aplicación A La Predicción De Rangos De Precios De Teléfonos Móviles

Autor: Ignacio Fernández Sánchez-Pascuala,
Tutor: Juan Tinguaro Rodríguez

GRADO DE MATEMÁTICAS. FACULTAD DE CIENCIAS MATEMÁTICAS.
UNIVERSIDAD COMPLUTENSE DE MADRID

8 de Julio de 2023



- 1 Introducción
- 2 Preliminares
- 3 Técnicas de Aprendizaje
 - Árboles para clasificación
 - K-Nearest-Neighbours
 - Redes Neuronales Artificiales
 - Regresión Logística Ordinal
- 4 Estudio Computacional
- 5 Conclusión



Contenido

1 Introducción

2 Preliminares

3 Técnicas de Aprendizaje

- Árboles para clasificación
- K-Nearest-Neighbours
- Redes Neuronales Artificiales
- Regresión Logística Ordinal

4 Estudio Computacional

5 Conclusión



Introducción

Los principales objetivos de este proyecto son:

- Introducir el concepto de aprendizaje automático. Conocer su evolución histórica, su contexto actual, sus aplicaciones y los diferentes tipos de aprendizaje automático.
- Detallar todos los pasos y procesos necesarios para el correcto funcionamiento de un algoritmo de aprendizaje supervisado.
- Explicar algunas de las técnicas de aprendizaje supervisado para clasificación y sus posibles modelos en función del objetivo del estudio a realizar.
- Adaptar los datos y técnicas de aprendizaje según el interés de un estudio en particular.
- Predecir el rango de precio de un móvil a partir de sus características mediante los algoritmos presentados en el trabajo.
- Comparar el rendimiento de los algoritmos introducidos en el estudio realizado.



Contenido

1 Introducción

2 Preliminares

3 Técnicas de Aprendizaje

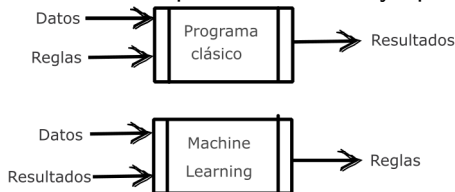
- Árboles para clasificación
- K-Nearest-Neighbours
- Redes Neuronales Artificiales
- Regresión Logística Ordinal

4 Estudio Computacional

5 Conclusión



- ¿Qué es el aprendizaje automático? Un poco de historia y aplicaciones.



- Tipos de aprendizaje automático: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.
- Aprendizaje para clasificación supervisada:
 $T = \{(x_1, y_1), \dots, (x_n, y_n)\}$, $x_i = (x_{i1}, \dots, x_{ip})$ representan las variables explicativas del dato i , e y_i representa su variable respuesta, $\forall i \in \{1, \dots, n\}$.

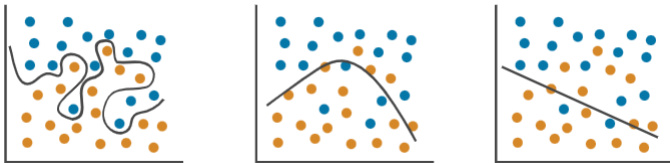
$$f : X \rightarrow Y = \{1, \dots, L\}$$

$$x \mapsto f(x) = \hat{y}$$



Preliminares

- Preprocesos: recopilación, preparación y limpieza de los datos.
- Generalización a nuevos datos: sobreajuste y subajuste, complejidad del modelo y separación de los datos.



$$E_i = (E_{i1} + E_{i2})/2, \quad E = \frac{E_1 + E_2 + E_3 + E_4 + E_5}{5} \quad (5 \times 2 \text{CV})$$

$$E = \frac{\sum_{i=1}^K E_i}{K} \quad (\text{K-Fold CV})$$



Modelo	Train 1	Train 2	...	Train k	Resultado
1	CV_{11}	CV_{21}	...	CV_{k1}	$\overline{CV_1}$
2	CV_{12}	CV_{22}	...	CV_{k2}	$\overline{CV_2}$
...
n	CV_{1n}	CV_{2n}	...	CV_{kn}	$\overline{CV_n}$

- Métricas de evaluación:

	Predicción positiva	Predicción negativa
Real positivo	TP	FN
Real negativo	FP	TN

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad Re = \frac{TP}{TP + FN}, \quad Pr = \frac{TP}{TP + FP}, \quad F_1 = \frac{2 \cdot Pr \cdot Re}{Pr + Re}$$



$$Re^M = \frac{1}{L} \sum_{i=1}^L Re_i, \quad Pr^M = \frac{1}{L} \sum_{i=1}^L Pr_i, \quad F_1^M = \frac{2 \cdot Pr^M \cdot Re^M}{Pr^M + Re^M} \quad (\text{Macro})$$

$$Re^\mu = \frac{\sum_{i=1}^L TP_i}{\sum_{i=1}^L (TP_i + FN_i)}, \quad Pr^\mu = \frac{\sum_{i=1}^L TP_i}{\sum_{i=1}^L (TP_i + FP_i)}, \quad F_1^\mu = \frac{2 \cdot Pr^\mu \cdot Re^\mu}{Pr^\mu + Re^\mu} \quad (\text{Micro})$$

$$MDP = \frac{\sum_{i=1}^m |y_i - \hat{y}_i|}{m}$$

- Estudio a realizar: Predicción rango de precio en teléfonos móviles



Contenido

1 Introducción

2 Preliminares

3 Técnicas de Aprendizaje

- Árboles para clasificación
- K-Nearest-Neighbours
- Redes Neuronales Artificiales
- Regresión Logística Ordinal

4 Estudio Computacional

5 Conclusión



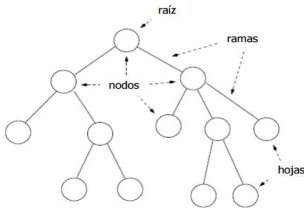
Contenido

- 1 Introducción
- 2 Preliminares
- 3 Técnicas de Aprendizaje
 - Árboles para clasificación
 - K-Nearest-Neighbours
 - Redes Neuronales Artificiales
 - Regresión Logística Ordinal
- 4 Estudio Computacional
- 5 Conclusión



Árboles para clasificación

- ¿En qué se basan? Estructura



- Algoritmo CART
- ¿Número de divisiones en un nodo? Factor de ramificación B
- ¿Qué test realizar en cada nodo?

$$i(T) = - \sum_{j=1}^L P(j) \log_2 P(j) \quad (\text{entropía de la información})$$



$$i(T, X_j) = \sum_{b \in B_{X_j}} P(b) \cdot i(T_b)$$

Se elige X_t tal que maximice la reducción de impureza en su crecimiento:

$$\begin{aligned}\Delta(T, X_j) &= i(T) - i(T, X_j) \\ X_t &= \arg \max_{X_j} \Delta(T, X_j)\end{aligned}$$

- Parada: Hiperparámetros β y m .



Árboles para clasificación

- Poda: Coste-Complejidad

$$E = \begin{cases} \frac{e+1}{N+|T|} & \text{si el árbol es una hoja} \\ \sum_{b \in B} P(b) \cdot E_b & \text{si no es una hoja} \end{cases} \quad (\text{error ponderado árbol})$$

Se selecciona el árbol podado que minimice el valor de E_α :

$$E_\alpha = E + \alpha \cdot K$$

siendo K el número de hojas del árbol y α hiperparámetro.

[Richard O. Duda,] [Kubat, 2017]



Contenido

- 1 Introducción
- 2 Preliminares
- 3 Técnicas de Aprendizaje
 - Árboles para clasificación
 - **K-Nearest-Neighbours**
 - Redes Neuronales Artificiales
 - Regresión Logística Ordinal
- 4 Estudio Computacional
- 5 Conclusión



K-Nearest-Neighbours

- Búsqueda k ejemplos más similares , x^j con $j \in \mathbb{K} = \{1, \dots, k\}$
- Similitud entre observaciones:

$$d(\mathbf{x}_i, \mathbf{z}_i) = \begin{cases} (\mathbf{x}_i - \mathbf{z}_i)^2 & \text{v.cuantitativa o v.ordinal} \\ B_{\mathbf{x}_i, \mathbf{z}_i} & \text{v.categórica} \end{cases}$$

$$d(\mathbf{x}, \mathbf{z}) = \sqrt{\sum_{i=1}^p d(\mathbf{x}_i, \mathbf{z}_i)}$$

[Kubat, 2017] [Jerome Friedman, 2008]



K-Nearest-Neighbours

- Normalización de variables:

$$x'_{ij} = \frac{x_{ij} - MIN_j}{MAX_j - MIN_j} \quad (\text{escalamiento min-max})$$

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad (\text{estandarización})$$

- Elección hiperparámetro k: Validación Cruzada.
- Mecanismo de votación:

$$\hat{\mathbf{y}} = f(\mathbf{x}) = \arg \max_{y \in \mathbb{Y}} \sum_{j \in \mathbb{K}} \delta_{y,y^j} \quad (\text{votación uniforme})$$

$$\hat{\mathbf{y}} = f(\mathbf{x}) = \arg \max_{y \in \mathbb{Y}} \sum_{j \in \mathbb{K}} w_j \delta_{y,y^j} \quad (\text{votación ponderada})$$

$$\text{donde } w_j = \begin{cases} \frac{d_k - d_j}{d_k - d_1} & d_1 \neq d_k \\ 1 & d_1 = d_k \end{cases}$$



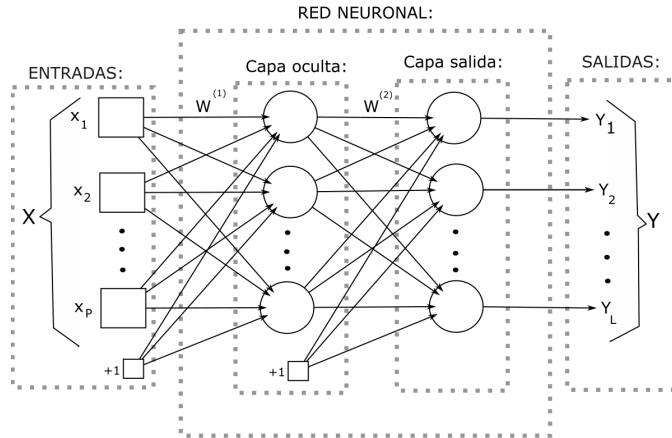
Contenido

- 1 Introducción
- 2 Preliminares
- 3 Técnicas de Aprendizaje**
 - Árboles para clasificación
 - K-Nearest-Neighbours
 - Redes Neuronales Artificiales**
 - Regresión Logística Ordinal
- 4 Estudio Computacional
- 5 Conclusión



Redes Neuronales Artificiales

- Estructura:



- Forward Propagation:

$$red_j^{(1)} = \sum_{i=0}^p x_i w_{ji}^{(1)} \equiv (\mathbf{w}_j^{(1)})^t \mathbf{x} \quad j \in \{1, \dots, n_H\}$$

$$h_j = \sigma(red_j^{(1)})$$

$$\sigma(v) = \frac{1}{1 + e^{-v}} \quad (\text{función de transformación})$$

$$red_k^{(2)} = \sum_{j=0}^{n_H} h_j w_{kj}^{(2)} \equiv (\mathbf{w}_k^{(2)})^t \mathbf{h} \quad k \in \{1, \dots, L\}$$

$$y_k = \sigma(red_k^{(2)})$$

$$g_k(\mathbf{x}) \equiv y_k = \sigma\left(\sum_{j=1}^{n_H} w_{kj}^{(2)} \sigma\left(\sum_{i=1}^p x_i w_{ji}^{(1)} + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$



- BackPropagation:

$$R(\mathbf{w}) \equiv - \sum_{i=1}^N \sum_{k=1}^L t_k(x_i) \log g_k(x_i) \quad (\text{función de entropía cruzada})$$

$$\Delta \mathbf{w} = -\eta \frac{\partial R}{\partial \mathbf{w}} \quad (\eta \text{ tasa de aprendizaje})$$

$$\Delta w_{mn} = -\eta \frac{\partial R}{\partial w_{mn}}$$

$$\mathbf{w}(m+1) = \mathbf{w}(m) + \Delta \mathbf{w}(m) \quad (\text{actualización pesos para cada iteracción})$$

$$\Delta w_{kj} = \eta \delta_k h_j \quad \text{con } \delta_k \equiv -\partial R / \partial \text{red}_k$$

$$\Delta w_{ji} = \eta \delta_j x_i \quad \text{con } \delta_j \equiv \sigma'(\text{red}_j) \sum_{k=1}^L w_{kj} \delta_k$$



- Entrenamiento: Algoritmo de Batch
- Técnicas para mejorar el Backpropagation: tratamiento previo variables explicativas, inicialización de los pesos, tasa de aprendizaje, número de neuronas, regularización.

$$R_{ef}(\mathbf{w}) = R(\mathbf{w}) + \frac{\alpha}{2} \mathbf{w}^t \mathbf{w}$$

[Richard O. Duda,] [Jerome Friedman, 2008] [Kubat, 2017]



Contenido

- 1 Introducción
- 2 Preliminares
- 3 Técnicas de Aprendizaje
 - Árboles para clasificación
 - K-Nearest-Neighbours
 - Redes Neuronales Artificiales
 - Regresión Logística Ordinal
- 4 Estudio Computacional
- 5 Conclusión



- Modelo:

$$P(Y \leq j | x) = \pi_1(x) + \cdots + \pi_j(x), \quad j \in \{1, \dots, L\} \quad \pi_j(x) \text{ probabilidad } x \text{ en la clase } j$$

$$P(Y \leq 1|x) \leq P(Y \leq 2|x) \leq \cdots \leq P(Y \leq L|x) = \sum_{j=1}^L \pi_j(x) = 1$$

$$L_j = \text{logit}[P(Y \leq j | x)] = \log \left(\frac{P(Y \leq j)}{1 - P(Y \leq j)} \right) = \log \left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_L} \right), j \in \{1, \dots, L-1\}$$

$$L_j = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_p x_p, \quad j \in \{1, \dots, L-1\}$$

$$\pi_j(x) = \frac{e^{\beta_{0j} + \sum_{k=1}^p \beta_k x_k}}{1 + e^{\beta_{0j} + \sum_{k=1}^p \beta_k x_k}} - \frac{e^{\beta_{0j-1} + \sum_{k=1}^p \beta_k x_k}}{1 + e^{\beta_{0j-1} + \sum_{k=1}^p \beta_k x_k}}, \quad j \in \{2, \dots, L-1\}$$

$$\hat{y} = f(x) = \arg \max_j \pi_j(x)$$



- Estimación de parámetros:

$$L(\beta) = \sum_{i=1}^N \sum_{j=1}^L y_{ji} \ln(\pi_j(x_i)) \quad (\text{función log-verosimilitud})$$

$$\beta(m+1) = \beta(m) - \left(\frac{\partial^2 L(\beta)}{\partial \beta^2} \right)^{-1} \frac{\partial L(\beta)}{\partial \beta} \quad (\text{actualización por Newton-Raphson})$$

- Significancia estadística del modelo: Test de Wald

$$H_0 : \hat{\beta}_k = 0$$

$$H_1 : \hat{\beta}_k \neq 0$$

$$W_k = \frac{\hat{\beta}_k}{\hat{SE}(\hat{\beta}_k)} \quad k \in \{1, \dots, p\}$$



Regresión Logística Ordinal

- Multicolinealidad
- Bondad de Ajuste: Hosmer-Lemeshow test
- Interpretación del modelo: Odds ratio

$$\hat{OR}_k = e^{\hat{\beta}_k}, \quad k \in \{1, \dots, p\}$$

- Tratamiento previo variables explicativas

[Jajang et al., 2022] [Agresti, 2007] [Fagerland and Hosmer, 2016]

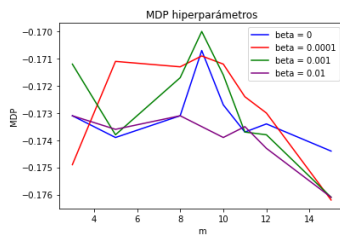
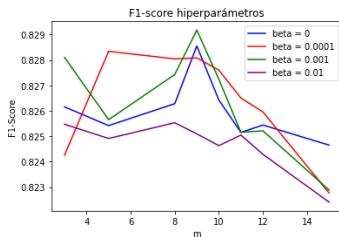
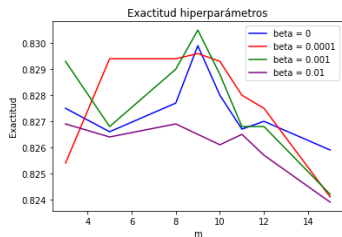


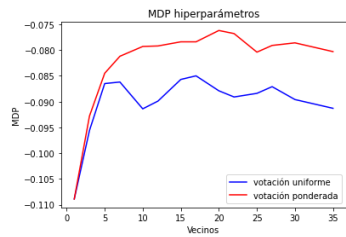
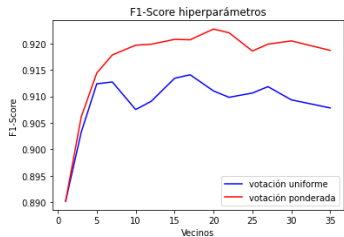
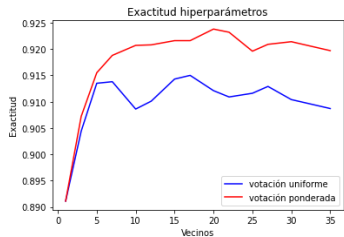
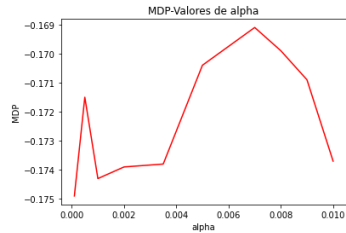
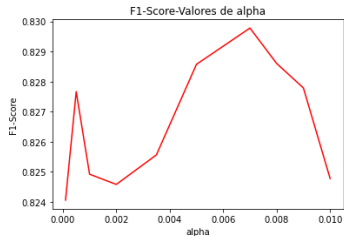
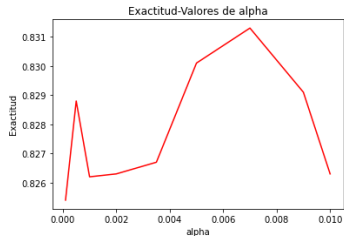
Contenido

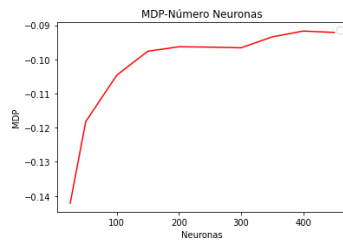
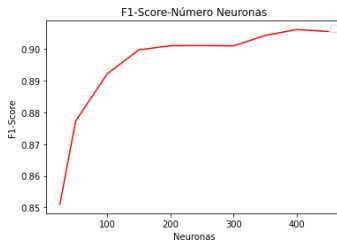
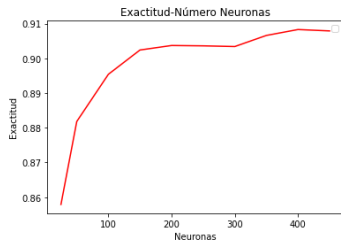
- 1 Introducción
- 2 Preliminares
- 3 Técnicas de Aprendizaje
 - Árboles para clasificación
 - K-Nearest-Neighbours
 - Redes Neuronales Artificiales
 - Regresión Logística Ordinal
- 4 Estudio Computacional**
- 5 Conclusión



- Conjunto de datos
- Configuración experimental
- Elección modelos:





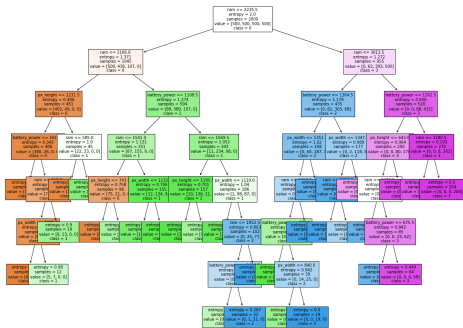


● Comparación resultados:

Modelos Finales	Exactitud	Macro F1-Score	MDP
Árbol parada	0.83	0.83	0.16
Árbol poda	0.84	0.84	0.15
k-NN	0.92	0.92	0.07
Red Neuronal	0.93	0.93	0.06
Reg.Log.Ordinal	0.98	0.98	0.02



Estudio Computacional



ram: 0.781234274768333

battery_power: 0.10878301796643593

px_width: 0.06597903968486062

px_height: 0.04400366758037049



Coeficientes

battery_power 0.032393

mobile_wt -0.072205

px_height 0.019161

px_width 0.018766

ram 0.052302

0/1 127.390004

1/2 3.889689

2/3 3.884086

Odds ratio

battery_power 1.032924e+00

mobile_wt 9.303403e-01

px_height 1.019346e+00

px_width 1.018943e+00

ram 1.053694e+00

0/1 2.112398e+55

1/2 4.889565e+01

2/3 4.862250e+01



Contenido

- 1 Introducción
- 2 Preliminares
- 3 Técnicas de Aprendizaje
 - Árboles para clasificación
 - K-Nearest-Neighbours
 - Redes Neuronales Artificiales
 - Regresión Logística Ordinal
- 4 Estudio Computacional
- 5 Conclusión







Conclusión

- Cumplimiento objetivos
- Investigaciones futuras



Referencias I

-  Agresti, A. (2007).
An Introduction to Categorical Data Analysis, volume 2nd Ed.
Wiley-Interscience.
-  Fagerland, M. W. and Hosmer, D. W. (2016).
Tests for goodness of fit in ordinal logistic regression models.
Journal of Statistical Computation and Simulation, 86(17):3398–3418.
-  Jajang, J., Nurhayati, N., and Mufida, S. J. (2022).
Ordinal logistic regression model and classification tree on ordinal response data.
BAREKENG: Jurnal Ilmu Matematika dan Terapan, 16(1):075–082.
-  Jerome Friedman, Robert Tibshirani, T. H. (2008).
The Elements of Statistical Learning, volume 2nd Ed.
Springer International Publishing AG.



Referencias II



Kubat, M. (2017).

An Introduction to Machine Learning, volume 2nd Ed.

Springer International Publishing AG.



Richard O. Duda, Peter E.Hart, D. G.

Pattern Classification, volume 2nd Ed.

