



DEPARTAMENTO DE MATEMÁTICA
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

Dígitos escritos a mano

Aplicaciones de la Matemática en Ingeniería

Clemente Ferrer
Cristian Marín
Gabriel Riffo

GRUPO 1

1 Reconocimiento de dígitos escritos a mano y sus aplicaciones

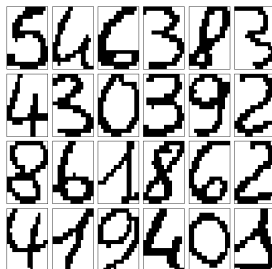
2 Resultados

- Limpieza de datos
- Software y librerías
- Indicadores de evaluación de los modelos
- Implementación de los clasificadores- KNN
- Análisis de los clasificadores

3 Conclusiones y análisis futuro

Reconocimiento de dígitos escritos a mano

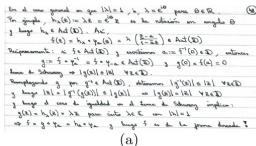
- Capacidad de las computadoras para reconocer manuscritos humanos.



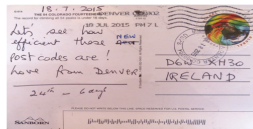
- **Objetivo:** Clasificar, a través de diversos métodos estudiados a lo largo del curso, dígitos escritos a mano según el número al cual representan.
- **Dataset:** Semeion Handwritten Digit, *UCI Machine Learning Repository*.

Aplicaciones

- (a) Digitalización de apuntes tomados en una clase.
- (b) Reconocimiento de matriculas de los automóviles.
- (c) Automatizar la redirección de cartas en el correo postal.
- (d) Procesamiento de cheques bancarios.



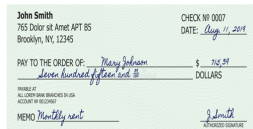
(a)



(c)



(b)



(d)

Limpieza de datos

1593 datos, cada uno con

- 1 256 atributos binarios.
- 2 10 posibles etiquetas.

1	2	3	4	5	6	7	8	9	10	...	257	258	259	260	261	262	263	264	265	266
0	0	0	0	0	1	1	1	1	1	...	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	1	...	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
0	0	0	1	1	1	1	1	1	1	...	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0

Dataset original.

Limpieza de datos

1593 datos, cada uno con:

- 1 256 atributos binarios.
- 2 1 etiqueta.

1	2	3	4	5	6	7	8	9	10	...	248	249	250	251	252	253	254	255	256	digit
1	1	1	1	1	1	1	1	1	1	...	0	0	0	0	0	1	1	1	0	9
0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	1
0	0	0	0	1	1	1	1	1	1	...	1	1	1	0	0	0	0	0	0	5
0	0	0	1	1	1	1	1	1	1	...	1	1	1	1	0	0	0	0	0	8
0	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	0	1
0	0	1	1	1	1	1	1	1	1	...	1	1	0	0	0	0	0	0	0	7
0	0	0	1	1	1	1	1	1	1	...	1	1	1	1	1	0	0	0	0	8

Dataset modificado.

Se utilizaron las siguientes librerías en R.

Clasificador	Librería	Función
Vecinos cercanos	class	knn
Support vector machine	kernlab	train/svmRadial
Regresión logística	nnet	multinom
Naive Bayes	e1071	naiveBayes
Árboles de decisión	rpart	mpart

Librerías estéticas: ggpubr, ggplot2, lattice, scales, formattable y gridExtra.

Precisión y coeficiente kappa

- Problema: Matriz de confusión 10x10.

Predicted \ Actual	0	1	2	3	4	5	6	7	8	9
9	1									7
8	3	2	6	1				1	11	2
7		1	1			1		8		
6	2		3		2		10	1	3	
5	1	1	1		4	16		1		6
4		1	1		9			1		
3	1	1	2	10		2			1	2
2		2	1							
1		6		1				2	1	1
0	15			1	3					

Ejemplo matriz de confusión.

- Precision: Numero de aciertos/Total de datos de testeo.

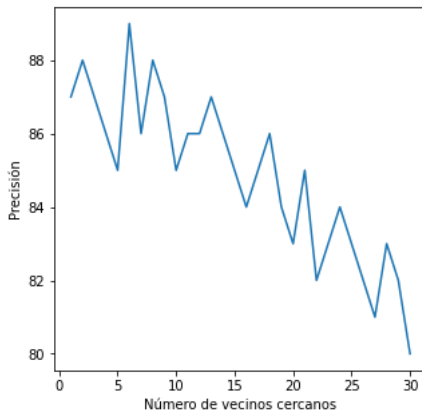
- Coeficiente kappa:

$$\kappa = \frac{N \sum_{i=1}^r x_{ii} - \sum_{i=1}^r x_{i+} \cdot x_{+i}}{N^2 - \sum_{i=1}^r x_{i+} \cdot x_{+i}},$$

- N : número de observaciones en la matriz.
- x_{ij} : el valor en la entrada ij de la matriz de confusión.
- x_{i+} : total de observación en la fila i .
- x_{+i} : total de observaciones en la fila en la columna i .
- r : el número de filas de la matriz.

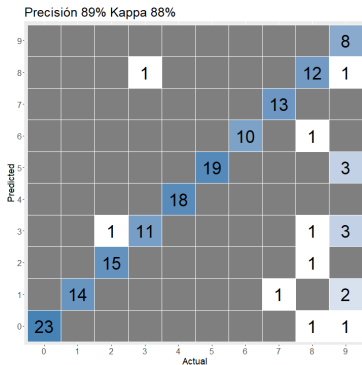
Implementación de los clasificadores

- KNN: ¿Cuál es el mejor número de vecinos?

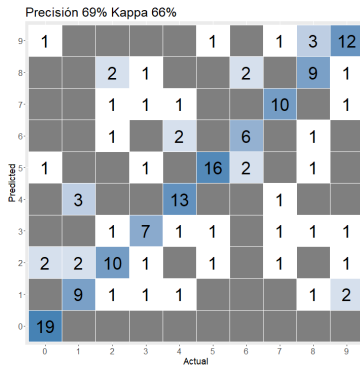


Análisis de los clasificadores

Se utilizó la matriz de confusión para obtener los indicadores mencionados.

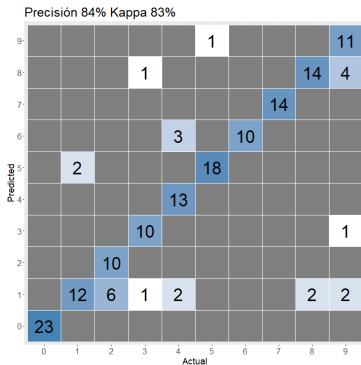


Vecinos cercanos, $k = 6$.

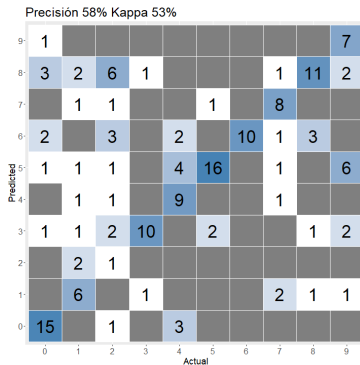


Regresión logística.

Análisis de los clasificadores



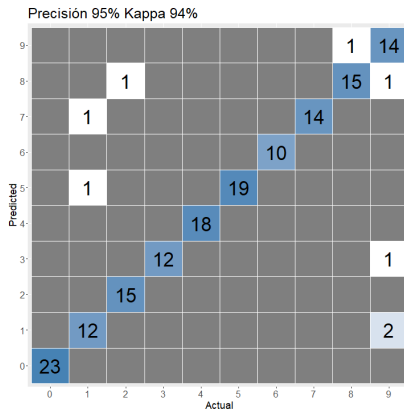
Naive Bayes.



Árboles de decisión.

Análisis de los clasificadores

El mejor clasificador según la literatura asociada:



Support vector machine.

Lo anterior se resume en la siguiente tabla:

Clasificador	Precisión (%)	κ(%)
Vecinos cercanos	89	88
Support vector machine	95	94
Regresión logística	69	66
Naive Bayes	84	83
Árboles de decisión	58	53

① Mejores y peores métodos

② Dificultades

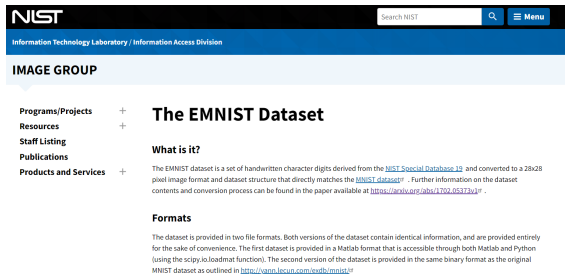
- ① Buscar otra manera de cuantificar la calidad de los clasificadores.
- ② Imposibilidad de la confección de una curva ROC.

El código fue subido a un repositorio público  MAT281.

Para complementar el estudio realizado podría considerarse lo siguiente:

- Dataset con más observaciones.
- Mejor resolución de los manuscritos.

Ejemplo: *National Institute of Standards and Technology* (NIST) que posee 280 000 observaciones.



The screenshot shows the NIST website's 'IMAGE GROUP' section. On the left is a sidebar with links: Programs/Projects, Resources, Staff Listing, Publications, and Products and Services. The main content area is titled 'The EMNIST Dataset' and includes a 'What is it?' section describing the dataset as handwritten character digits from the NIST Special Database 19, converted to a 28x28 pixel format. It also mentions a 'Formats' section, noting that the dataset is provided in both Matlab and Python (scipy.io.loadmat) formats. The page has a search bar and a menu icon at the top right.

- Ben-David, Arie. About the relationship between ROC curves and Cohen's kappa. *Eng. Appl. of AI*. 21. 874-882. 2008.
- Ahamed, Hafiz & Alam, Ishraq & Islam, Md. SVM Based Real Time Hand-Written Digit Recognition System. 2019
- Colas, Fabrice & Brazdil, Pavel. Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. *International Federation for Information Processing Digital Library; Artificial Intelligence in Theory and Practice*. 217. 2007.
- Landgrebe, Thomas & Duin, Robert. Approximating the multiclass ROC by pairwise analysis. *Pattern Recognition Letters*. 28. 1747-1758. 2017.
- Cohen, Gregory, et al. EMNIST: Extending MNIST to handwritten letters. 2017 *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017.



DEPARTAMENTO DE MATEMÁTICA
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

Dígitos escritos a mano

Aplicaciones de la Matemática en Ingeniería

Clemente Ferrer
Cristian Marín
Gabriel Rizzo

GRUPO 1