Análisis del rendimiento del algoritmo de clasificación k-nn

Sergio Marrero Marrero

Máster Universitario en Sistemas Inteligentes y Aplicaciones Numéricas en la Ingeniería(MUSIANI)
Asignatura: Metodología de la I+D y Documentación Científica
Universidad Las Palmas de Gran Canaria(ULPGC)

5 de julio de 2016

Resumen

Se realiza en este documento un estudio sobre el rendimiento del algoritmo de clasificación *K-nn*. Se estudia el rendimiento del algoritmo trabajando en distintas condiciones. Se prueban dos métodos diferentes para calcular la distancia: la distancia Euclídea y la distancia de Manhattan. El otro parámetro que se varía es el número de vecinos, de este se escogen diez valores diferentes, concretamente los números impares contenidos en el intervalo $(0, 20)$. De cada combinación de los factores alterados se 20 obtienen réplicas, cada una de ellas con un conjunto de entrenamiento y test diferente. Una vez se obtienen los resultados se emplean diversas herramientas de inferencia estadística para consolidar las distintas conclusiones que se extraen de este experimento.

1. Introducción

El contexto de este documento es la asignatura *Metodología de la I+D y Documentación Científica* del *Máster Oficial de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería (SIANI)(año académico 2015/2016)*. Se ha realizado como tarea final para superar la asignatura, el diseño de un experimento consistente en analizar el rendimiento del clasificador *K-vecinos más cercanos (k-nn)*. Este proyecto fue llevado a cabo con el lenguaje de programación R [6]. Además, cabe resaltar que a parte de las diapositivas expuestas en clase se ha seguido el libro de fundamentos de estadística: Estadística aplicada básica [5].

El desarrollo de este documento se estructura de la siguiente forma, en la sección 2 se examinan los datos que se van a utilizar. En la sección 3 se define la metodología del experimento, detallando así aspectos relevantes relativos a los parámetros y métodos que se van a utilizar, así como los pasos que se van a seguir para llevar a cabo el experimento. En la sección 4 se

realizará un análisis de la interacción entre los factores, tanto visual como númerico (ANOVA) con la intención de sacar a la luz que factores intervienen en los distintos resultados.En esta sección se comprueba también si hay evidencias estadísticas para afirmar que los dos mejores resultados obtenidos pertenecen a poblaciones diferentes. Finalmente, en las sección 5 se detallan las conclusiones obtenidas.

2. Base de datos: Sistema de radares para la clasificación de la ionosphera

El conjunto de datos que se van a analizar lleva como título *Johns Hopkins University Ionosphere database* y procede de un estudio realizado por el *Space Physics Group of The Johns Hopkins University Applied Physics Laboratory*[8]. En este apartado se dará dar una pequeña descripción de los datos, lo necesario para poder comprender los resultados. Para una información detallada de cómo se obtuvo dicha base de datos se puede consultar [2] y [3]. El estudio (del cual resultaron los datos en cuestión) tenía como objetivo analizar las características físicas de la ionosfera. Con la finalidad de estudiar las características de la emisión y recepción de las ondas de radio, se situó un sistema de radares de alta frecuencia en la localidad de Goose Bay, Labrador. Dependiendo de las características físicas de la ionosfera, estas ondas de radio rebotarían o atravesarían dicha capa. Por otro lado, la variable que caracterizará el experimento será una función compleja de autocorrelación (ACF), la cual depende de la forma en que la señal emitida es recibida.

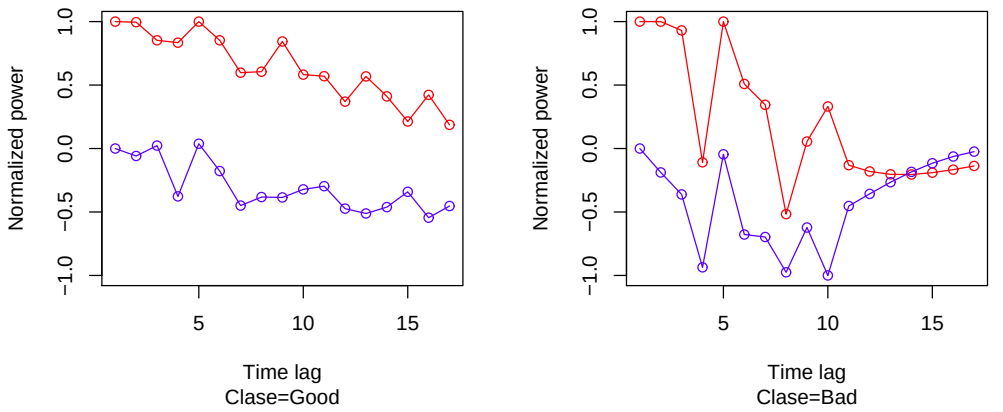
En este experimento, la ACF estará descrita por 17 valores, cada uno de ellos relacionado con la potencia electromagnética recibida en 17 pulsos dentro de un intervalo de tiempo. Cada uno de estos valores está compuesto por parte real y parte imaginaria, resultando 34 valores diferentes. Estos 34 valores constituyen los 34 primeros atributos o columnas de la base de datos en cuestión. La columna 35 constituye la clase de las muestras, dividiéndose en dos valores nominales: *buena recepción (g)* y *mala recepción (b)*. Hay un total de 351 instancias de las cuales 126 pertenecen a la clase *g* y 225 a la clase *b*. No se registra ningún valor perdido.La tabla 1 muestra un subconjunto de las 6 primeras filas del conjunto de datos en cuestión. Debido al gran número de columnas sólo se ha mostrado las 4 primeras y las tres últimas columnas. Todos los atributos contienen datos numéricos, exceptuando a la columna 35 (la clase) que contiene datos nominales, precisamente el carácter *g (good)* y *b (bad)* haciendo referencia respectivamente a *buena recepción* y *mala recepción*. Estos 34 valores servirán como entradas para el clasificador. Todos los valores han sido normalizados al rango $[-1, 1]$.

Instancia	Real1	Imag1	Real2	Imag2	...	...	Real17	Imag17	Class
1	1	0	1.00	-0.06	...	...	0.19	-0.45	g
2	1	0	1.00	-0.19	...	...	-0.14	-0.02	b
3	1	0	1.00	-0.03	...	...	0.56	-0.38	g
4	1	0	1.00	-0.45	...	...	-0.32	1.00	b
5	1	0	1.00	-0.02	...	...	-0.05	-0.66	g
6	1	0	0.02	-0.01	...	...	-0.00	0.12	b

Tabla 1: Muestra de las 6 primeras filas del conjunto de datos *Johns Hopkins University Ionosphere database*

En la Figura 1 se puede observar los datos de las dos primeras filas de la tabla 1. El eje

x contiene las distintas columnas (lapsos de tiempo o *times lag*) y en la ordenada se han representado sus valores (*ACF normalizados*). La imagen de la izquierda 1a representa a la clase g y la de la derecha 1b a la clase b.



(a) Primera fila del conjunto de datos (b) Segunda fila del conjunto de datos

Figura 1: Representacion de las dos primeras filas del conjunto de datos. Las imagen izquierda pertenece a la clase g y la imagen derecha a la clase b. El color rojo representa a la parte real. El color azul representa la parte imaginaria

3. Metodología

3.1. Algoritmo k-nn

El método de clasificación que se utilizará en el experimento es el *k-nn* (por sus siglas en ingles *k nearest neighboor*). Este es un método de clasificación supervisada no paramétrico que estima o predice el valor de la clase de una muestra a partir de la clase de sus k vecinos.

El método se desarrolla brevemente en el Algoritmo 1. El conjunto de valores $D = \{(\mathbf{x}_1, c_1), \mathbf{x}_2, c_2), \ldots (\mathbf{x}$ es el subconjunto de muestras clasificadas, en donde x_i es la muestra o instancia i, la cual se compone a su vez de $x_1^i, x_2^i, \ldots, x_n^i$ componentes o atributos y c_i es la clase a la que pertenece dicha muestra. La muestra que se desea clasificar es $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Para mas información sobre este método mirar en la bibliografía [7].

Algorithm 1 K vecinos más cercanos

procedure K-NN$(D = \{(\mathbf{x}_1, c_1), \mathbf{x}_2, c_2), \ldots (\mathbf{x}_N, c_N)\}, \mathbf{x} = (x_1, x_2, \ldots, x_n))$
  for *todo objeto ya clasificado* $(\mathbf{x}_i, c_i)$ **do**
    calcular $d(i) \longleftarrow d(\mathbf{x}_i, \mathbf{x})$ ▷ Generamos vector de distancias
  end for
  Sort(**d**) ▷ Ordenamos d en forma ascendente o descendente
  Seleccionamos los K objetos $\mathbf{x_i}$ mas cercanos a $\mathbf{x}$
  Asignamos a $\mathbf{x}$ la clase mas frecuente de los K más cercanos
end procedure

3.2. Diseño y planificación del experimento

Se desea comprobar la efectividad del algoritmo funcionando con distintos parámetros. La efectividad del algoritmo será medida a través de la medida de rendimiento *accuracy (ACC)* 1, la cual se define como el cociente entre el número de muestras correctamente clasificadas y el número total de muestras clasificadas. El cociente de dicha ecuación se compone de muestras de la clase positiva correctamente clasificadas TP y las muestras de la clase negativa TF correctamente clasificadas y en el denominador se encuentra el tamaño total de muestras que se clasificaron N.

$$ACC = \frac{TP + TN}{N} \tag{1}$$

Los parámetros del algoritmo que se quieren controlar son los siguientes. Por un lado se quiere comparar el rendimiento del algoritmo trabajando con dos funciones de distancia diferentes, y por otro lado se quiere comprobar su rendimiento utilizando diferente número de vecinos. Se utilizó como referente para las distintas distancias, la distancia de Minkowski Ecuación 2. la cual, particularizada para $q = 1$ y $q = 2$ constituyen la distancia de Manhattan y Euclidea respectivamente Ecuación 3 y Ecuación 4. Por otro lado, el número de vecinos que se utilizó fue $K = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$.

$$d_{Mink} = \left(\sum_{k=1}^{n} \mid x_{ik} - x_{jk} \mid^{q} \right)^{\frac{1}{q}} \tag{2}$$

$$d_{Manh} = \sum_{k=1}^{n} \mid x_{ik} - x_{jk} \mid \tag{3}$$

$$d_{Euc} = \sqrt[2]{\sum_{k=1}^{n} \mid x_{ik} - x_{jk} \mid^{2}} \tag{4}$$

Se resume de manera pormenorizada los distintos factores de los que depende el experimento:

- Variable respuesta: *Accuracy(ACC)*
- Factor A: Tipo de distancia. Los niveles de este factor son dos:
  - A_1 : Euclídea
  - A_2 : Manhattan
- Factor B: Número de vecinos. Los niveles son:
  - $B_1 : K = 1$
  - $B_2 : K = 3$
  - ...
  - $B_i : K = impar(i)$

- ...
- $B_{10} : K = 19$

donde cada una de las combinaciones A_iB_j se corresponde con un tipo de tratamiento diferente.

.

El diseño del experimento responde a un diseño factorial de 20 réplicas por cada tratamiento La tabla 2 resume las características del experimento.

	A_1	A_2
B_1	$y_{1,1,1}, y_{1,1,2}, ..., y_{1,1,k} ... y_{1,1,20}$	$y_{1,2,1}, y_{1,2,2}, ..., y_{1,2,k} ... y_{1,2,20}$
B_2	$y_{2,1,1}, y_{2,1,2}, ..., y_{2,1,k} ... y_{2,1,20}$	$y_{2,2,1}, y_{2,2,2}, ..., y_{2,2,k} ... y_{2,2,20}$
B_3	$y_{3,1,1}, y_{3,1,2}, ..., y_{3,1,k} ... y_{3,1,20}$	$y_{3,2,1}, y_{3,2,2}, ..., y_{3,2,k} ... y_{3,2,20}$
...	...	...
B_{10}	$y_{10,1,1}, y_{10,1,2}, ..., y_{10,1,k} ... y_{10,1,20}$	$y_{10,2,1}, y_{10,2,2}, ..., y_{10,2,k} ... y_{10,2,20}$

Tabla 2: Diseño factorial del experimento. Veinte réplicas en cada tratamiento

Por otro lado, la técnica que se siguió para estimar el comportamiento del modelo fue la de *hold-out*. El *hold-out* consiste en separar el conjunto de datos en una proporción destinada a conformar el clasificador o subconjunto *train* y la otra proporción destinada a realizar el test. En este caso las proporciones fueron del 60 % y 40 % respectivamente para *train* y para *test*. Debido a que la clase no se distribuye de forma homogénea sobre el conjunto de datos total, el *hold-out* se realizó de manera estratificada proporcional, garantizando una representación proporcional en ambos subconjuntos de datos.

Se realizó un *hold-out* cada vez que se llevó a cabo el experimento, garantizando así que cada réplica representa es obtenida con un subconjunto *train* y *test* procedente de un muestreo diferente. El algoritmo 2 muestra de manera conceptual los pasos seguidos para realizar el experimento.

Para realizar la clasificación con el algoritmo *K-nn* se utilizó la función *knnVCN()* del paquete *knnGarden* [9], para realizar el *holdout()* se usó la función *holdout* del paquete [1].

Algorithm 2 Diseño del experimento

```
1:  procedure CLASIFICACIONKNN(data)                ▷ data: Conjunto de datos
2:      for j ={Euclidea, Manhattan} do             ▷ Factor Distancia
3:          for i ={1, 3, 5, 7, 9, 11, 13, 15, 17, 19} do   ▷ Factor Vecinos
4:              set(seed)                           ▷ Iniciamos semilla
5:              for k= {1, 2, 3 ... 19, 20} do      ▷ Réplicas
6:                  seed++                          ▷ Cambiamos la semilla
7:                  (train_ijk, test_ijk) ⟵HoldOut(data)   ▷ Hacemos holdout estratificado
8:                  ACC_ijk) ⟵ K-NN(train_ijk, test_ijk)   ▷ Estimamos la clase con K-NN
9:                  save(ACC_ijk)                   ▷ Almacenamos la respuesta
10:             end for
11:         end for
12:     end for
13: end procedure
```

4. Resultados

4.1. Test de normalidad: test de Jarque-Bera

En primer lugar, antes de plantear cualquier tipo de hipótesis se deberá asumir que los resultados siguen algún tipo de distribución de probabilidad. En base a las observaciones de los histogramas, se plantea la hipótesis de que siguen una distribución normal. Por tanto, la hipótesis nula será asumir que cada tratamiento A_iB_j contiene una muestra de resultados procedenetes de una distribución normal $N(\mu_{ij}, \sigma_{ij})$, la hipótesis alternativa será que no siguen esta distribución. El contraste de hipótesis se realizará con un test de Jarque-Bera

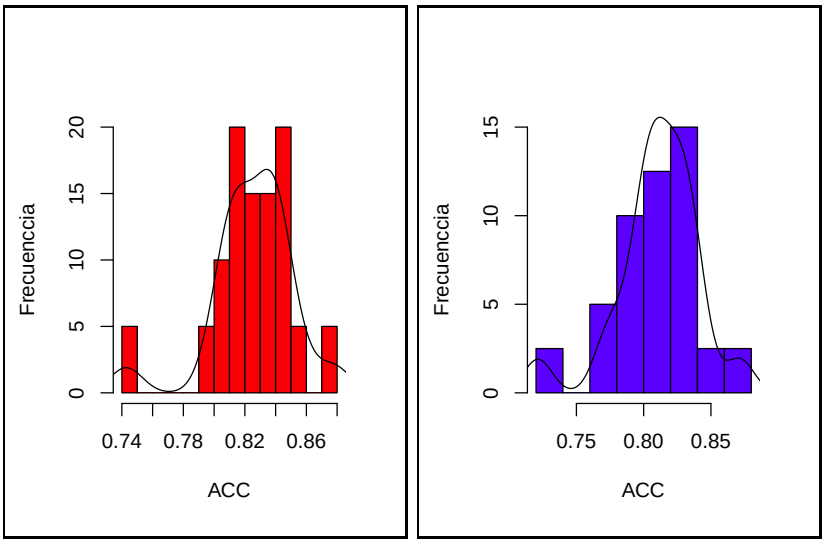
Se utilizará para discriminar un valor de significación de $\alpha = 0{,}05$. Como el test que se realizará será de dos colas, habrá que tener cuenta el valor medio de α . La tabla 3 muestra los peores valores obtenidos en dicho test. La primera fila contiene el resultado del tratamiento A_1B_8, es decir, factor distancia al nivel *Euclidea* combinada con 15 vecinos y la segunda fila contiene el resultado del tratamiento A_1B_{10}, es decir, factor distancia al nivel *Euclidea* combinada con 19 vecinos.

Para realizar el test de Jarque-Bera se uso la función *jarque.test()* del paquete *moments* [4].

	JB	p-Value	H_0
A_1B_8	9,883	$0{,}007714 < \frac{\alpha}{2}$	No se acepta
A_1B_{10}	5,95	$0{,}051 > \frac{\alpha}{2}$	Se acepta

Tabla 3: Test de normalidad de Jarque-Bera

El primero de ellos no pasa el test. El segundo si lo pasa, pero queda muy cercano al valor umbral $\alpha/2$. Se procede a continuación a revisar sus histogramas. En la figura 2a y 2b se puede observar los histogramas junto la estimación de la curva de densidad de estos tratamientos, respectivamente. Se observa en estas imágenes que existen razones para pensar que el test de Jarque-Bera esta siendo condicionado por posibles outliers y porque la lista de réplicas es relativamente pequeña. Debido a que el resto de los tratamientos pasaron el test con un *valor-P* significativamente grande, y debido también a que los histogramas muestran una distribución aparentemente normal, se asumirá que la distribución de errores del tratamiento A_1B_{15} proviene de una normal.



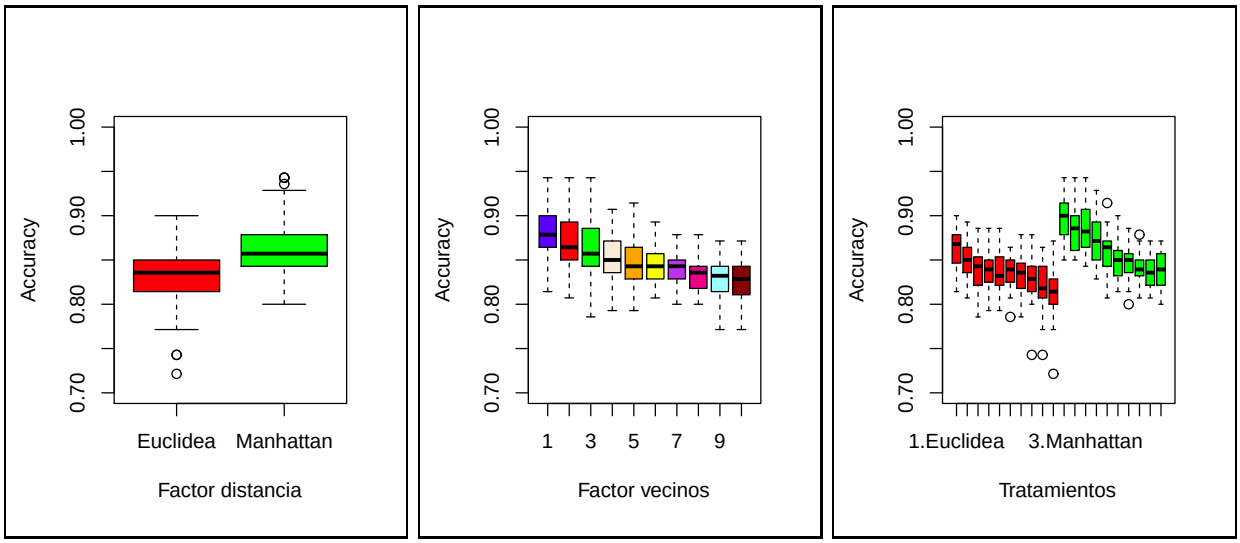
(a) Tratamiento: A_1B_{15} (b) Tratamiento:A_1B_{19}

Figura 2: Histogramas de los tratamientos con menor valor-P en el test de normalidad de Jarque-Bera

4.2. Interacción entre los factores. Análisis de la varianza

4.2.1. Análisis visual

Las tres imágenes de la Figura 3 muestran una representación en diagrama de cajas de los distintos factores y tratamientos frente a la variable respuesta. La primera imagen por la izquierda 3a muestra el diagrama de cajas para el factor *distancia*. En este se aprecia cierta relación entre la respuesta y el tipo de distancia utilizada, es decir, parece existir relación entre ambos factores. En la imagen central 3b se representa el diagrama de cajas para el factor *vecinos*. También parece indicar una relación entre ambos factores, en este caso inversa. La tercera imagen 3c presenta el diagrama de cajas con los distintos tratamientos. Esta imagen refuerza la idea de que los factores A y B están relacionados con la respuesta, pero no hay ninguna indicación de que ambos factores interactúen entre sí.



(a) Factor: Distancia (b) Factor: Número de vecinos (c) Factor: Tratamientos

Figura 3: Diagrama de cajas de los distintos factores y tratamientos

4.2.2. Análisis de la varianza de dos factores. ANOVA

A continuación se pasa a realizar un análisis de la varianza. Para realizar el ANOVA se utilizó la función *aov()* del paquete estadístico *stats*, el cual forma parte de los paquetes por defecto de R.

Las hipótesis de partida son las siguientes:

- $H_0^A : \tau_1 = \tau_2$
- $H_1^A : \tau_1 \neq \tau_2$
- $H_0^B : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_{10}$
- $H_1^B : \beta_i \neq \beta_j$
- $H_0^{AB} : (\tau\beta)_{11} = (\tau\beta)_{12} = (\tau\beta)_{21} = (\tau\beta)_{22} = (\tau\beta)_{31} = (\tau\beta)_{32} = \ldots = (\tau\beta)_{10,10}$
- $H_1^{AB} : (\tau\beta)_{ij} \neq (\tau\beta)_{ij}$

tal que τ_i representa el efecto del nivel i del factor A, β_j representa el efecto del nivel j del factor B y $(\tau\beta)_{ij}$ representa el efecto conjunto de la interacción de ambos factores.

Los resultados se muestran en la Tabla 4. La primera columna de la tabla contiene a los distintas fuentes de interacción, la segunda contiene los grados de libertad de cada fuente de interacción, la tercera columna contiene la suma de los cuadrados, la cuarta columna contiene los cuadrados medios. Estas tres últimas columnas nos permiten calcular el estadístico F que se encuentra en la quinta columna. La sexta columna contiene el *valor P* asociado a dicho valor F.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
B	1	0.11	0.11	197.58	2×10^{-16}
A	1	0.07	0.07	113.83	2×10^{-16}
AB	1	0.00	0.00	7.18	0.0077
Residuals	396	0.23	0.00		

Tabla 4: Resultados del análisis de varianza(ANOVA) de dos factores

Si utilizamos como significación de $\alpha = 0{,}05$, se pueden asumir las siguiente conclusiones:

- Siendo el *valor-P* de: $2 \times 10^{-16} < \frac{\alpha}{2}$, se debe rechazar la hipótesis H_0^B y aceptar H_1^B. Es decir, existe influencia entre el factor *número de vecinos* y la variable respuesta. Es bastante evidente la influencia de dicho factor sobre la variable respuesta, ya que el *valor P* es insignificante.
- Siendo el *valor-P* de: $2 \times 10^{-16} < \frac{\alpha}{2}$, se debe rechazar la hipótesis H_0^A y aceptar H_1^A. Es decir, existe influencia entre el factor *tipo de distancia* y la variable respuesta.En este caso también es bastante evidente la influencia de dicho factor sobre la variable respuesta, ya que el *valor P* es también insignificante.
- Siendo el *valor-P* de: $0{,}0077 < \frac{\alpha}{2}$, se debe rechazar la hipótesis H_0^{AB} y aceptar H_1^{AB} Es decir, los dos factores de control interactúan entre sí. Esta propiedad característica no fue percibida analizando los datos visualmente.

4.3. Comparación de los mejores resultados: t-Student

A continuación se comparan las mejores combinaciones que se obtuvieron para cada nivel del factor *distancia*. Para realizar el test *t-Student* se utilizó la función *t.test()* del paquete estadístio *stats*. Las mejores combinaciones de los factores son las mostradas en la tabla 5:

	$\overline{x}$	s_p
A_1B_1	0,86	0,02
A_2B_1	0,89	0,02

Tabla 5: Media muestral y error estandar de los tratamientos A_1B_1 y A_2B_1

A continuación se va a comprobar que dichos valores no son fruto de la casualidad, sino que realmente pertenecen a dos poblaciones con medias poblacionales diferentes. Para ello se realizará un *test de hipótesis de Student.* Las hipótesis son las siguientes:

- $H_0 : \mu_{A_1B_1} = \mu_{A_2B_1}$
- $H_1 : \mu_{A_2B_1} > \mu_{A_1B_1}$

con lo cual el test se tendrá que realizar de forma unilaterar, hacia la derecha. El resultado de este test se muestra en la tabla 6. La primera columna muestra el estadístico t, la segunda columna los grados de libertad, la tercera columna el $p-Value$, la cuarta columna el intervalo en el que se estima donde se encuentra la media de los errores y finalmente en la última columna se resalta qué hipótesis se acepta. A la vista de los resultados la hipótesis nula debe ser rechazada

y se debe aceptar que la el tratamiento A_2B_1 pertenece a una población cuya media es superior a la del tratamiento A_1B_1, tal y como se sospechaba.

t	df	p-Value	Intervalo	H_0
$-4{,}4878$	38	$3{,}2 \times 10^{-05}$	$[0{,}0214, INF]$	No se acepta

Tabla 6: Contraste de hipótesis *t-Student* para los mejores resultados de lo tratamientos

5. Conclusiones

Se planteó como objetivo en este documento comprobar la efectividad del clasificador *K-nn* variando dos de sus parámetros, la distancia: distancia *Euclidea* y distancia de *Manhattan*, y por otro lado el número de vecinos, utilizando para ello 10 niveles correspondientes al número de vecinos $k = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$.

Se pudo comprobar a través de un análisis de la varianza (ANOVA) que el rendimiento del algoritmo (variable respuesta) estaba condicionado por el factor *distancia*, por el factor *número de vecinos* y por la interacción de estos dos factores. Se demostró que la mejor combinación de los parámetros para este algoritmo se consigue utilizando la distancia de Manhattan y un sólo vecino más cercano. Además, independientemente de la distancia utilizada, el aumento del número de vecinos provocó en todas las situaciones un empeoramiento de los resultados. No se debe de terminar sin decir que hubo un tratamiento cuya normalidad fue cuestionada por el *test de Jarque-Bera* ver 4.1, sin embargo, despúes de ver su histograma, se consideró que este hecho pudo ser causado por la poca cantidad de réplicas y la existencia de posibles outliers, y por ello se decidió presuponer la normalidad.

Como líneas futuras se propone comprobar el rendimiento del algoritmo *k-nn* usando otras distancias y otro número de vecinos. También se propone probar otro tipo de algoritmos de clasificación, como por ejemplo *Support Vector Machine (SVM)* con el objetivo de comprobar si se puede mejorar el *ACC*.

6. Agradecimientos

Como colofón de este trabajo, cabe agradecer al *Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería* por haber permitido y habilitado los recursos necesarios para realizar este trabajo. A los profesores Javier Lorenzo Navarro y Luis Alberto Padrón Hernández por haber guiado cada paso de esta investigación.

Referencias

[1] Paulo Cortez. *rminer: Data Mining Classification and Regression Methods*, 2015. R package version 1.4.1.

[2] R.A. Greenwaltd, K.B. Baker, R. A. Hutchins, and C. Hanuise. And hf phased array radar for studying small-scale structure in the hugh latitude ionosphere. *Radio Science*, (20):63–79, 1985.

[3] R.A. Greenwaltd, K.B. Baker, J. P. Villain, and S. Wing. Spectral characteristcs oh high frequency backscatter from high latitude ionospheric irregularities: A statistical survey. Technical report, Rome Air Development Center, Marth 1987.

[4] Lukasz Komsta and Frederick Novomestky. *moments: Moments, cumulants, skewness, kurtosis and related tests*, 2015. R package version 0.14.

[5] David S. Moore. *Estadística aplicada básica*. 1995.

[6] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[7] B. W. Silverman and M. C. Jones. E. fix and j.l. hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951). *International Statistical Review / Revue Internationale de Statistique*, 57(3):233–238, 1989.

[8] Space Physics Group of The Johns Hopkins University Applied Physics Laboratory. Johns hopkins university ionosphere database, 1989.

[9] Boxian Wei, Fan Yang, Xinmiao Wang, and Yanni Ge. *knnGarden: Multi-distance based k-Nearest Neighbors*, 2012. R package version 1.0.1.