
Algoritmo kNN - Javier E. Camarillo Polanco

Anonymous Author(s)

Affiliation

Address

email

1 Introducción

- 2 El algoritmo kNN (por sus siglas en inglés “k nearest neighbors”) es un algoritmo de reconocimiento
3 de patrones el cual consiste en la clasificación de un elemento no conocido mediante la obtención de
4 la moda de clases de los ‘ k ’ elementos más cercanos. Siendo “ k ” una constante mayor que 0.

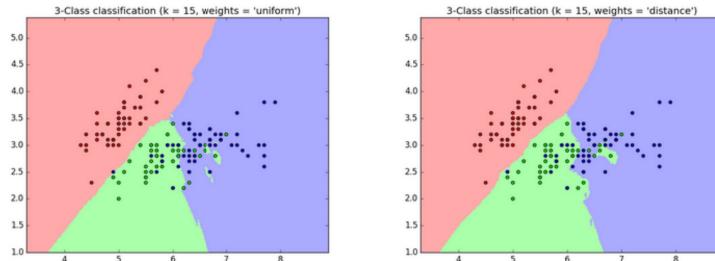


Figure 1: Ejemplo de kNN

- 5 En muchos casos la aplicación del algoritmo incluye la graficación de los elementos en n dimensiones,
6 y la distancia entre ellos se obtiene mediante la aplicación de diferentes distancias, con la euclidiana
7 siendo la más común.

8 2 Aplicaciones

- 9 Las aplicaciones pueden abarcar dominios desde el entretenimiento hasta la medicina.

10 2.1 Ejemplo 1

- 11 La aplicación popular de películas “Netflix” emplea el uso del algoritmo de clasificación kNN
12 para etiquetar el género de la película basado en propiedades aparentemente arbitrarias (tales como
13 cantidad de explosiones o besos) para determinar la categoría de la película.

14 2.2 Ejemplo 2

- 15 En medicina se pueden registrar los diferentes síntomas de un paciente para detectar la enfermedad
16 potencial que pueda padecer el paciente (la clase, en este caso, sería la enfermedad)

17 **3 Implementación del algoritmo kNN en un examen Vocacional**

18 En este proyecto seguiremos la realización de un test de tipo vocacional para estudiantes de la
19 Universidad Autónoma de Chihuahua, Facultad de Ingeniería categorizando a alumnos con perfiles
20 de personalidad mediante encuestas en carreras de:

21 Ingeniería Aeroespacial, Ingeniería Civil/Topografía, Ingeniería en Sistemas Hardware/ Ciencias de
22 la computación, Ingeniería Físico-Matemática.

23 **4 Hipótesis**

24 Según perfiles de personalidades, las diferencias son más pronunciadas en estudiantes de diferentes
25 áreas, en lugar de diferentes licenciaturas de ingeniería, por lo cual un algoritmo de vocaciones tipo
26 ingeniería puede no ser tan preciso como un test orientado a las diferentes carreras que abarcan el
27 espectro completo de personalidades, como normalmente se realizan.

28 **5 Encuestas**

29 Las encuestas se realizan con preguntas de personalidad generalmente orientadas a preferencias
30 personales para hacerlo más entretenido al usuario, así como pocas preguntas relacionadas a las
31 aspiraciones profesionales del estudiante.

32 **5.1 Recopilación de datos (Aplicación de encuestas)**

33 Para aplicar las encuestas se recopilaron 40 encuestas de Ingeniería Civil y Topografía, 30 de
34 Ingeniería aeroespacial y 30 de Ingeniería en Sistemas Hardware, como bonus 20 encuestas de
35 estudiantes Físico Matemáticos, pero el algoritmo solo tratará con la relación entre las primeras 3
36 áreas.

Aeroespacial	Pop	Salgo con mis amigas a ver Fast and Furious	Perros	Cerveza	Verde
Aeroespacial	Rock	Juego videojuegos Star Wars	Perros	Café	Verde
Aeroespacial	Pop	Hago deporte Star Wars	Ninguno	Café	Rojo
Aeroespacial	Banda	Salgo con mis amigos Avengers	Perros	Cerveza	Rojo
Aeroespacial	Pop	Hago deporte Avengers	Perros	Café	Rojo
Aeroespacial	Pop	Juego videojuegos Fast and Furious	Gatos	Café	Rojo
Aeroespacial	Pop	Salgo con mis amigos Star Wars	Perros	Cerveza	Azul
Aeroespacial	Rock	Juego videojuegos Star Wars	Perros	Café	Verde
Aeroespacial	Pop	Juego videojuegos Avengers	Perros	Cerveza	Rojo
Aeroespacial	Rock	Salgo con mis amigos Star Wars	Perros	Café	Azul
Aeroespacial	Pop	Salgo con mis amigos Avengers	Ninguno	Cerveza	Verde
Aeroespacial	Rock	Hago deporte Avengers	Perros	Cerveza	Azul
Aeroespacial	Banda	Salgo con mis amigos Avengers	Perros	Alcohol	Azul
Aeroespacial	Rock	Juego videojuegos Avengers	Perros	Café	Verde
Civil/Topografía	Pop	Salgo con mis amigos Avengers	Perros	Cerveza	Rojo
Civil/Topografía	Rock	Hago deporte Avengers	Perros	Café	Azul
Civil/Topografía	Pop	Salgo con mis amigos Fast and Furious	Perros	Cerveza	Azul
Civil/Topografía	Banda	Salgo con mis amigos Fast and Furious	Perros	Café	Azul
Civil/Topografía	Rock	Salgo con mis amigos Star Wars	Perros	Cerveza	Verde
Civil/Topografía	Banda	Salgo con mis amigos Fast and Furious	Ninguno	Café	Rojo
Civil/Topografía	Pop	Salgo con mis amigos Avengers	Perros	Café	Azul
Civil/Topografía	Rock	Salgo con mis amigos Fast and Furious	Perros	Cerveza	Azul
Civil/Topografía	Banda	Salgo con mis amigos Star Wars	Perros	Cerveza	Azul
Civil/Topografía	Rock	Salgo con mis amigos Avengers	Perros	Cerveza	Azul

Figure 2: Algunas de las preguntas son sobre películas populares o colores favoritos

37 **5.2 Conversión de preguntas a datos numéricos**

38 Para convertirlos a variables separadas por comas simplemente utilizamos una herramienta llamada
39 “Unit Conversion” para reemplazar las preguntas de cada encuesta por identificadores 1,2 y 3
40 respectivamente, así como 4 variables en la primera columna a las cuales les corresponden las

41 etiquetas Ae, Cv, Cc y Fm para Aeroespacial, Civil, Ciencias de la computación y Físico-Matemático
42 respectivamente.

```
Ae,1,2,3,3,2,1,3,2,1,2,1,1,1,1,2,3,1,2  
Ae,2,1,2,1,1,3,2,3,1,2,2,1,3,3,3,1,1,3,3,1  
Ae,1,2,1,1,2,3,1,3,2,2,1,1,3,1,3,1,3,3,2  
Ae,2,3,1,3,2,1,1,2,1,2,3,1,3,1,1,1,2,1,3,1  
Ae,3,1,3,1,1,1,3,2,3,2,1,1,1,1,3,1,2,1,3,2  
Ae,2,3,3,1,2,1,3,2,1,2,3,2,3,1,1,1,3,3,1  
Ae,2,2,2,3,2,1,1,2,1,3,2,2,3,1,1,1,2,1,3,1  
Ae,2,1,1,1,1,2,2,2,3,1,1,2,1,1,1,3,1,3,2  
Ae,1,2,1,1,2,3,2,3,3,2,1,3,1,3,1,2,1,1,2  
Ae,2,2,3,1,1,1,3,3,1,2,1,2,1,1,1,3,3,1  
Ae,1,1,1,1,2,2,3,1,2,1,1,1,1,1,3,3,3,2  
Ae,2,1,3,3,1,3,2,2,2,3,1,3,1,1,1,2,3,3,1  
Ae,1,3,3,1,1,2,2,2,2,1,2,3,1,1,1,3,3,1,3  
Ae,3,1,3,1,3,2,1,2,2,2,1,1,1,3,1,1,2,3,3,2  
Ae,1,2,3,1,2,3,2,2,2,1,2,3,1,1,1,3,1,1,3,2  
Cv,2,1,3,1,1,1,3,2,1,2,2,1,1,2,1,1,1,1,0,2  
Cv,1,3,3,1,2,2,3,3,2,2,2,1,3,2,3,1,2,1,3,2  
Cv,2,1,2,1,1,2,2,1,1,2,2,3,1,1,3,1,2,1,1,2  
Cv,3,1,2,1,2,2,2,1,3,2,2,1,3,1,1,1,1,1,1,2  
Cv,1,1,1,1,1,3,3,2,2,2,1,1,3,1,3,1,2,3,1,2
```

Figure 3: Encuesta convertida a un archivo de variables separadas por comas

43 6 Programación en Python 3

44 Las siguientes funciones se pueden consultar de los archivos python en el link de github adjunto.

45 Para realizar el archivo “kNN 4.0” en python, primero importamos la biblioteca ‘csv’ mediante la cual
46 se adjuntan valores de [0,0,0],[1,0,0],[0,1,0] y [0,0,1] para cada variable 0,1,2 y 3 respectivamente a
47 la lista que genera un ciclo en main.

48 La principal razón de convertir las variables en este dicótomo binario consiste en que las respuestas
49 entre las preguntas deben de tener la misma distancia, ya que, en su mayoría, una pregunta de tipo:
50 ¿Qué color te gusta? implica una respuesta única cuya distancia es igual respecto a las otras 2.

51 Generando así una lista de 20 listas por pregunta en la encuesta con vectores de 3 elementos, es
52 decir una matriz de 3 por 20 por “tamaño”, siendo tamaño una variable definida por el tamaño de la
53 encuesta previamente medida por la siguiente función:

54 $tamano = medir_{enc}()$

55 Enseguida tenemos una lista

56 $sur_v = list()$

57 A la cual se le adjunta un objeto de clase “survey” y a cada uno, una encuesta contestada como
58 atributo “ans”.

59 Finalmente un ciclo adjunta un atributo “cat” (corto de categoría) a cada objeto de clase “survey” de
60 la lista.

61 6.1 Clasificación

62 Para el proceso de clasificación, se adjuntan con un ciclo las distancias definidas por la función
63 $get_{dist}(v1, v2)$, cuyos argumentos son la encuesta a clasificar y la encuesta con la que se está
64 comparando, la cual también se le asigna a un atributo “dist” de distancia, a cada objeto “survey”.

65 Finalmente, el valor de “k” se le pide al usuario y con un insertion sort se ordenan las distancias
66 de o a “k” de la lista, las cuales se ordenan al comparar cada distancia con las distancias ordenadas

67 de la lista y la categoria (cat) de la encuesta a clasificar es igual a la moda del vector de categorías
68 ordenadas de 0 a k y se imprime la coincidencia del vector en la consola.

69 **6.2 Instalar los módulos para graficar los datos**

70 Introducimos "import matplotlib.pyplot as plt" en el encabezado
71 para importar la librería con la cual graficar los datos, la
72 librería matplotlib se obtiene para python 3 en windows, llamando
73 a la consola en la carpeta de C:Users "Usuario"AppData Local Programs Python o en donde esté
74 ubicado python en su ordenador,
75 dando shift click derecho en la misma y ejecutando el comando:
76 "pip install"
77 Para instalar un administrador de paquetes de python llamado "pip" con el cual se instala el paquete
78 "matplotlib" con el comando:
79 "pip install matplotlib"
80 Para instalar pip en otras plataformas y para obtener más información puede visitar la página oficial
81 en el siguiente vínculo.
82 <https://pip.pypa.io/en/stable/installing/>
83 Y para más información con el módulo matplotlib puede visitar:
84 <https://matplotlib.org/>

85 **6.3 Representación Gráfica**

86 Se extraen cuatro colores en valor hexadecimal de <https://material.io/> y se guardan en variables,
87 ejemplo:
88 `color1 ='white'`
89 Se importa de matplotlib
90 `import matplotlib.patches as mpatches`
91 Para crear la leyenda, en este ejemplo se crea la etiqueta de "Aeroespacial" y se implementa en la
92 leyenda:
93 `labAe = mpatches.Patch(color = col1, label ='Aeroespacial')`
94 Lo mismo se hace para las otras cuatro clasificaciones.
95 Y se grafica la posición del vector originalmente no clasificado, cuya posición inicial es 0 +(distancia
96 0 -1), simplemente para propósitos estéticos y que no aparezca muy lejos de los otros datos con el
97 marcador "marker='x'" para representarlo con una cruz.
98 `plt.scatter(0 + kdist[0] - 1, 0, 100, color ='white', marker ='x')`
99 Los demás datos se grafican con las siguiente función con un ciclo, la cual la puede consultar en el
100 código fuente al igual que las demás.
101 `plot('Ae', col1, val * 16, 's', 1, 0)`
102 `plot('Cv', col2, val * 8, 's', 1, 0)`
103 `plot('Cc', col3, val * 4, 's', 1, 0)`
104 `plot('Fm', col4, val, 's', 1, 0)`

Table 1: kNN

Tasa de coincidencia			
kNN	Aeroespacial	Civil	Sistemas
1	50%	60%	30%
2	50%	90%	20%
3	40%	90%	30%
4	40%	90%	20%
5	40%	90%	20%
6	40%	90%	20%
7	40%	90%	10%

105 **6.4 Entrenamiento de datos**

106 Cómo saber la precisión del algoritmo, es decir, ¿aproximadamente en qué porcentaje de casos nos
 107 agrupará con personas de personalidad similar por carrera? La respuesta es mediante un proceso de
 108 entrenamiento que toma lugar en algoritmos de reconocimiento de patrones.

109 El proceso de entrenamiento toma lugar en el programa “kNN training.py”, el cual verifica la precisión
 110 de categorización del algoritmo kNN al ejecutar la misma secuencia de clasificación en 10 encuestas
 111 de cada una de las 3 categorías a medir, es decir: 10 encuestas de aeroespacial se comparan con
 112 toda la “caja” de encuestas, posteriormente; 10 encuestas de Sistemas y 10 de Ingeniería Civil se
 113 comparan de igual manera.

114 En este caso, al determinar la precisión de 10 encuestas, por cada encuesta que acierte su categoría
 115 original se añaden un 10% a su precisión, obteniendo los resultados anteriores.

116 Lo cual indica que posiblemente el mayor número de coincidencia se obtiene con $k = 4$ y $k = 3$, para
 117 balancear más la proporción y sacrificar 10% de aeroespacial para compensar la baja coincidencia de
 118 sistemas.

119 **7 Observaciones**

120 Típicamente el algoritmo kNN implica graficación an al menos 2 o 3 dimensiones mientras que
 121 éste solamente grafica en una, ya que, como se observó en una versión Beta del proyecto, el cual
 122 asignaba variables en 3 direcciones basado en sus preguntas, los datos coinciden arbitrariamente y su
 123 representación en la gráfica dependía de una simple coincidencia con la cantidad de unos dos y tres
 124 de las encuestas, lo cual no representaba la distancia real entre ellas, y sobreempalmaba los datos
 125 desproporcionadamente, por lo cual se eliminaron las coordenadas y y se le asignó la distancia pura a
 126 la coordenada x , por lo que los datos a graficar simplemente son la representación ordenada de los
 127 elementos a la encuesta arbitraria.

128 La pobre precisión al determinar la coincidencia con alumnos de Ciencias de la computación/ Sistemas
 129 Hardware refleja la diversidad en perfiles de personalidad entre los estudiantes de dichas carreras en
 130 la facultad, puesto que tienen una distancia demostrablemente grande entre ellas.

131 **Agradecimientos**

132 Este proyecto fue desarrollado bajo la guía del doctor Luis Carlos González Gurrola.

133 **Referencias**

134 [1] 9 algorithms that changed the future - John Mc.Cormick [2] Data Algorithms - Mahmoud Parsian [3] Machine
 135 learning for hackers - Drew Conway & John Miles White [4] Machine learning in action - Peter Harrington