Universidad de Talca
Facultad de Ingeniería
Departamento de Ciencias de la Computación

Tarea Nº 1 - KNN

Machine Learning

Gonzalo Mardones Baeza
gmardones10@alumnos.utalca.cl
9 de septiembre de 2017

1. Introducción

La presente tarea da a conocer el comportamiento del conjunto de datos **Iris**, el cual se compone de 150 observaciones de flores de iris de tres especies diferentes. Hay 4 mediciones de flores dadas: longitud del sépalo, anchura del sépalo, longitud del pétalo y el ancho del pétalo, todos en la misma unidad de centímetros. El atributo predicho es la especie, que es uno de setosa, versicolor o virginica. Se utilizará el algoritmo de clasificación K-NN(k-Nearest Neighbors), con el objetivo de realizar un clasificador de datos y generar un modelo para predecir la clase a la que pertenece cada registro del dataset.

2. Validación y Resultados Experimentales

Para el proceso experimental del proyecto, se dividió el conjunto iris que posee 150 instancias en 10 subconjuntos de 15 instancia cada uno. Para formar el conjunto de entrenamiento del modelo, se utilizaron nueve subconjuntos y un subconjunto para probarlo. Una vez leído estos valores desde el archivo CSV, se generó una lista de instancias que fueron reubicados aleatoriamente para obtener resultados distintos en cada ejecución del programa.

Se realizó un procedimiento de validación cruzada para las tareas de entrenamiento y de prueba, estas se repiten para cada uno de los subconjuntos. Para cada valor K, se inicia con el primer subconjunto como prueba y los nueve restantes para entrenamiento, una vez finalizado la evaluación se utiliza el segundo subconjunto para prueba, los restantes para entrenamiento y así sucesivamente.

Para evaluar el clasificador K-NN se emplearon valores de K entre 1 y 10

Se obtiene de la certeza (exactitud) para cada valor K por medio de la evaluación de todos los subconjunto de prueba y entrenamiento, obteniendo un valor de certeza promedio para cada valor K. Los resultados generados son variados en cada una de las pruebas realizadas, ya que en cada ejecución del programa se generaban listas distintas, aunque es posible mencionar que según el cuadro (1), se determinó que para 11 ejecuciones del programa los mejores resultados se encontraban con valores $K = 9$ y $K = 10$, mientras que los peores resultados fueron identificados para valores de $K = 1$, $K = 2$ y $K = 3$.

Resultados de K

K	Máx	Mín
1	0	11
2	0	8
3	0	5
4	3	2
5	4	1
6	3	0
7	5	0
8	4	1
9	7	2
10	9	0

Cuadro 1: *Se realizaron 11 pruebas, existieron casos que el valor máximo se encontró en $K = 10$ y $K = 9$, por lo que se marco como valor máximo en esas 2 pruebas, situación similar para el caso de los valores mínimos.*

3. Requisitos previos

Es necesario contar con el dataset de estudio y las tecnologías previamente configuradas, el dataset se encuentra en formato **CSV**, en el link:
`https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data`

El programa se desarrollo bajo el IDE **Sublime Text, Build 3126**, en el lenguaje de programación **Python** versión 2,7,13, S.O **macOS Sierra** versión 10,12,6, es necesario contar con los módulos **numpy,matplotlib** de Python que deben ser incluidos externamente, por medio de la sentencia en el terminal:

```
MP-GM:Tarea_knn gmardones $ sudo pip install numpy
MP-GM:Tarea_knn gmardones $ sudo pip install matplotlib
```

El programa es ejecutado por medio del terminal en el directorio del proyecto con la sentencia:

```
MP-GM:Tarea_knn gmardones $ python tarea_knn.py
```

Generando la siguiente salida desde el terminal y gráfica (ver Figura 1):

```
MP-GM:Tarea_knn gmardones$ python tarea_knn.py
k = 1, Exactitud: 96.0%           |  k = 6, Exactitud: 96.6666666667%
k = 2, Exactitud: 96.0%           |  k = 7, Exactitud: 96.6666666667%
k = 3, Exactitud: 96.0%           |  k = 8, Exactitud: 96.6666666667%
k = 4, Exactitud: 96.0%           |  k = 9, Exactitud: 97.3333333333%
k = 5, Exactitud: 96.6666666667% |  k = 10, Exactitud: 97.3333333333%
```

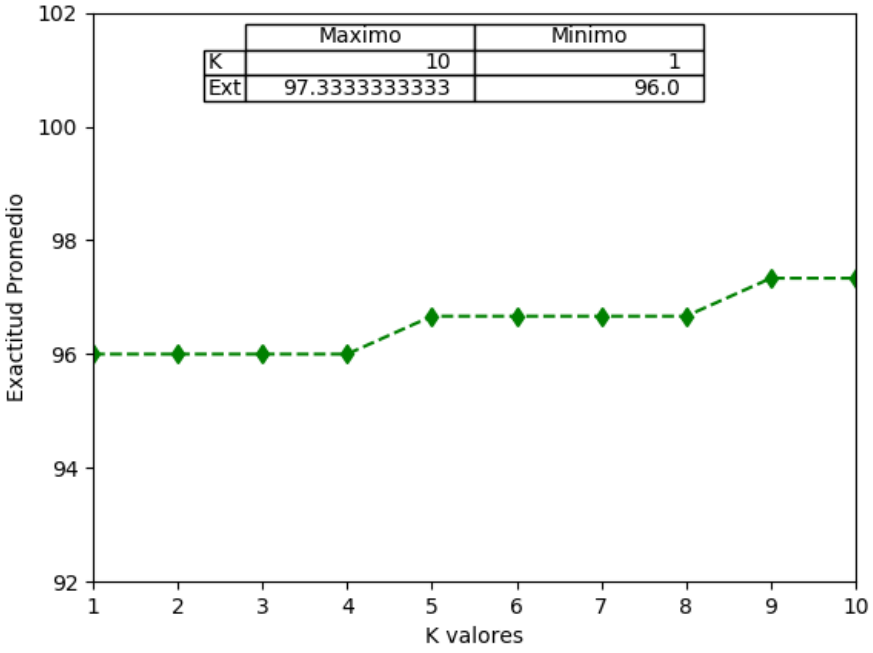

Figura 1: *Gráfico de línea, exactitud promedio para cada valor K* - Fuente propia.