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

Algoritmo KNN

Alumno: Miguel Arenas Santander
Profesor: Cesar Astudillo
Fecha: 15 de Septiembre del 2017.

1. Introducción

KNN (K-Nearest Neighbors) es un algoritmo fácil de entender e implementar, y una poderosa herramienta para tener a su disposición. El modelo para kNN es el dataset completo. Cuando se requiere una predicción para una instancia de datos no vista, el algoritmo kNN buscará en el dataset de entrenamiento para las k-similares instancias. El atributo de predicción de las instancias más similares se resume y devuelve como la predicción para la instancia invisible. La medida de similitud depende del tipo de datos. Para datos reales, se puede usar la distancia euclidiana. En el caso de problemas de regresión, puede devolverse el promedio del atributo predicho. En el caso de clasificación, la clase más prevalente puede ser devuelta.

2. Problema

El conjunto de datos es estándar donde la especie es conocida para todas las instancias. De esta manera, se dividirán los datos en conjuntos de datos de entrenamiento y prueba, para así utilizar los resultados para evaluar la implementación de nuestro algoritmo. *El problema con el cual trabajaremos es la clasificación del iris. Para esto, trabajaremos con 150 observaciones de flores de iris de 3 diferentes especies. Existen 4 medidas de las flores dadas (longitud del sépalo, ancho del sépalo, longitud del pétalo y ancho del pétalo). Lo que se desea predecir es la especie de la flor (setosa, virginica o versicolor) con los datos observador.*

El conjunto de datos es estándar donde la especie es conocida para todas las instancias. De esta manera, se dividirán los datos en conjuntos de datos de entrenamiento y prueba, para así utilizar los resultados para evaluar la implementación de nuestro algoritmo.

3. Solución

Se tiene un documento con los datos de las flores, estas se cargarán y dividirán en 2 tipos de datos, los de entrenamiento que serán el 90 % de estos y el 10 % restante serán para realizar las pruebas para verificar el funcionamiento del algoritmo. Con los datos de entrenamiento se calcularán las distancias euclidianas entre los almacenados para pruebas y los de entrenamiento. Estos serán etiquetados y ordenados de menos a mayor para así determinar los vecinos más cercanos. Una vez realizado esto, creamos una nueva lista, pero con las predicciones realizadas y luego los compararemos con los valores de pruebas.

4. Resultados

*Al ver una nula diferencia entre los distintos valores de K, se probó utilizando distintos valores para la validación cruzada del set de muestras, donde aquí se pudo observar una diferencia en el porcentaje de predicción de la variedad de iris tal como se presenta en la tabla.*Se realizaron exactamente 150 pruebas, en las cuales se evaluó 15 veces para cada valor de k, al observar la variedad de posibles resultados para cada uno de los valores de K, se pudo observar una cierta tendencia, en donde, el porcentaje variaba entre un 86,667 % como mínimo y un 100 % de certeza en la predicción. Estos valores (no siempre constantes) eran los resultados más comunes en conjunto con 93,333 %. Por lo cual, no se logró diferencia una clara diferencia entre los diferentes valores de K. Aun así, se logró observar que mientras mayor era la cantidad de vecinos existía una mayor probabilidad de predicción. Esto quiere decir, que se obtuvieron más valores cercanos al 100 % con mayores valores en K, mientras que, cuando era menor el valor de la variable la mayor parte de los resultados se aproximaba a un 86 %.

Al ver una nula diferencia entre los distintos valores de K, se probó utilizando distintos valores para la validación cruzada del set de muestras, donde aquí se pudo observar una diferencia en el porcentaje de predicción de la variedad de iris tal como se presenta en la tabla.

Cantidad de entrenamiento	Cantidad de prueba	% certeza predicción
135	15	entre 86,667 y 100 %
120	30	entre 83,333 y 100 %
100	50	entre 77,551 y 97,959 %
75	75	entre 78,333 y 97,333 %

Para cada una de estas pruebas se ejecutó 15 veces el script con un valor de 3 para la variable k, obteniendo múltiples valores en el porcentaje de certeza de la predicción. La tabla presenta el porcentaje mínimo y máximo en la predicción en la variedad de iris.

5. Conclusión

La predicción de situaciones se puede lograr utilizando técnicas de machine learning. El utilizar knn, uno de los algoritmos más sencillos en esta área y con gran probabilidad en la predicción, ayudo a comprender como funciona esta parte de la ciencia de la computación dando una leve idea de su potencial en el mercado laboral.