Predicción de la estructura secundaria de proteínas globulares

Maria Lucas

2023-03-24

Contents

Algoritmo k-NN

El algoritmo k-nn (k-nearest neighbors) es un algoritmo de aprendizaje supervisado utilizado para clasificación y regresión. En el proceso de clasificación, el algoritmo busca encontrar la clase más común entre los k ejemplos de entrenamiento más cercanos al punto de consulta. En el proceso de regresión, el algoritmo busca encontrar el valor medio de los k ejemplos de entrenamiento más cercanos al punto de consulta.

El funcionamiento del algoritmo k-nn es bastante sencillo. Primero, se carga un conjunto de datos de entrenamiento que consta de entradas y etiquetas correspondientes. Luego, se toma un punto de consulta (una entrada sin etiquetar) y se calcula la distancia entre ese punto y cada punto en el conjunto de datos de entrenamiento. Las distancias más comunes utilizadas en k-nn son la distancia euclidiana y la distancia Manhattan.

Una vez que se han calculado las distancias, se seleccionan los k puntos de entrenamiento más cercanos al punto de consulta. Si se está realizando clasificación, se seleccionan las etiquetas correspondientes a estos puntos de entrenamiento y se toma la etiqueta más común como la etiqueta asignada al punto de consulta. Si se está realizando regresión, se toma el valor medio de las etiquetas de los k puntos de entrenamiento más cercanos como el valor asignado al punto de consulta.

La elección del valor de k en el algoritmo k-nn es un factor crítico que puede afectar significativamente el rendimiento del modelo. Si k es demasiado pequeño, el modelo puede ser sensible al ruido en los datos y puede sobreajustarse. Si k es demasiado grande, el modelo puede subajustarse y no ser capaz de capturar patrones sutiles en los datos. El valor de k dependerá del conjunto de datos y el problema específico que se está abordando, aunque se puede empezar por la raíz cuadrada del número de datos e ir ajustando.

Ventajas	Inconvenientes
Simple y fácil de interpretar	No produce un modelo
Rápida fase de entrenamiento	Lenta fase de clasificación
No paramétrico	Se debe escoger una k apropiada
Buen rendimiento en datos con pocos atributos	Computacionalmente costoso para gran cantidad de datos
Se puede actualizar a tiempo real con nuevos datos	Sensible a datos redundantes y a valores atípicos
	Requiere pre-procesamiento de los datos

Codificación one-hot

```
print('Ni puta idea de como hacer esto hulio')
```

```
## Ni puta idea de como hacer esto hulio
```

Clasificador knn

noc

Carga del fichero

alfjpajdf

onehot