<!DOCTYPE html>

KNN para el análisis de variables cuantitativas (regresión)

KNN para el análisis de variables cuantitativas (regresión)

1 Predicción con un regresor

El objetivo es predecir los valores que tomará una variable cuantitativa conocidos los valores de otra u otras variables. Por ejemplo, se quiere predecir el peso de los pingüinos si se conoce la longitud de su aleta. El peso es la variable respuesta y la longitud de la aleta es el regresor (también se conoce como variable independiente, cofactor, predictor,...). Cuando la variable respuesta es cuantitativa se suele denominar problema de regresión.

En este caso se quiere predecir el peso de un pingüino cuya aleta mide 210 mm. Para ello se va a utilizar el algoritmo K-Nearest Neighbors (KNN). La idea es calcular la predicción como la media de los k-valores más cercanos a 210 mm. Primero se calcula la distancia al regresor que se va a predecir:

Ahora se ordenan los datos en función de la distancia al regresor predicho:

Ya se puede calcular la predicción:

Si K = 1 se utiliza el valor más cercano. Luego la predicción para el peso es 5050 g.

Si K = 2 la predicción es la media entre 5050 y 5300 = 5175 g.

Si K = 3 la predicción es la media entre 5050, 5300 y 4750 = 5033.3333333 g.

Y así para otros valores de K.

2 Predicción con m regresores

Se quiere predecir el peso de un pingüino con long_pico = 40 mm, prof_pico = 18 mm y long_aleta = 210 mm, es decir, se tiene la información de tres regresores. En el caso de que se tengan dos o más regresores el algoritmo sigue siendo el mismo: la predicción es la media de los k valores más cercanos. En general se utiliza la distancia euclídea para calcular la distancia entre cada observación y los datos a predecir.

$$x = (x_1, x_2, ..., x_m), \quad y = (y_1, y_2, ..., y_m)$$

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2}$$

Con la función anterior se calcula la distancia entre los regresores y el dato a predecir:

Ahora se ordenan los datos en función de la distancia al regresor predicho:

Si K = 1 la predicción para el peso es 5050 g.

Si K = 2 la predicción es la media entre 5050 y 5300 = 5175 g.

Si K = 3 la predicción es la media entre 5050, 5300 y 4750 = 5033.3333333 g.

Y así sucesivamente.

3 Normalización de los regresores

En el ejemplo anterios se han utilizado 3 variables para calcular la distancia:

long_pico, que toma valores en torno a 40 mm;

prof_pico, que toma valores alrededor de 20 mm;

long_aleta, con valores alrededor de 200 mm.

Por tanto, la longitud de la aleta va a tener más influencia en la magnitud de la distancia que las otras dos variables. En cierta manera se desperdicia la información aportada por long_pico y prof_pico. Por ello se

suelen normalizar los regresores, de manera que todos ellos aporten información en las mismas condiciones. La manera tradicional de reescalar variables en KNN es la normalización min-max. Este proceso transforma una variable cualquiera de manera que pasa a tomar valores en el rango 0-1. La fórmula que consigue dicha transformación es:

$$x = (x_1, x_2, ..., x_m) \Rightarrow x^* = (x_1^*, x_2^*, ..., x_m^*)$$

$$x_i^* = \frac{x_i - min(x)}{max(x) - min(x)}$$

Con esta transformación se consique que $min(x^*) = 0$ y $max(x^*) = 1$.

Se va a repetir el ejemplo anterior pero con las variables normalizadas entre 0 y 1. Primero se define la función de normalización:

Se normaliza:

También se tiene que normalizar el dato que se quiere predecir:

Y ahora se calculan las distancias y se ordena:

Si K = 1 la predicción para el peso es 3400 g.

Si K = 2 la predicción es la media entre 3400 y 5300 = 4350 g.

Si K = 3 la predicción es la media entre 3400, 5300 y 4250 = 4316.6666667 g.

4 Regresores cualitativos

La variable respuesta obligatoriamente tiene que ser cuantitativa ya que estamos en un problema de regresión. Hasta el momento sólo hemos utilizado regresores cuantitativos pero también pueden ser cualitativos. Por ejemplo, se va a introducir en el análisis el regresor especie. En este caso se quiere predecir el peso de un pingüino con long_pico = 40 mm, prof_pico = 18 mm, long_aleta = 210 mm y especie = Gentoo. Para poder aplicar el algoritmo KNN se necesita calcular la distancia entre los datos: ¿como se calcula la distancia entre Adelie, Gentoo y Chinstrap? La solución consiste en introducir variables auxiliares 0-1. Se van a crear tres variables auxiliares, una por nivel de la variable factor: Adelie1, Gentoo1 y Chinstrap1.

Adelie1 = 1 si especie = Adelie, y Adelie1 = 0 si especie $\neq$ Adelie.

Chinstrap1 = 1 si especie = Chinstrap, y Chinstrap1 = 0 si especie $\neq$ Chinstrap.

Gentoo1 = 1 si especie = Gentoo, y Gentoo1 = 0 si especie $\neq$ Gentoo.

Por tanto, el dato que se quiere predecir es:

Se crea la base de datos:

No es necesario escalar estas variables porque ya están en el rango 0-1. Se calculan las distancias y se ordena:

Si K = 1 la predicción para el peso es 5300 g.

Si K = 2 la predicción es la media entre 5300 y 5050 = 5175 g.

Si K = 3 la predicción es la media entre 5300, 5050 y 4750 = 5033.3333333 g.

5 Regresores cualitativos-2

En realidad no es necesario utilizar las tres variables auxiliares, con dos de ellas es suficiente. Por ejemplo, se va a utilizar Adelie1 y Chinstrap1:

Los pinguinos Adelie se indican como (Adelie1=1, Chinstrap1=0);

Los pinguinos Chinstrap como (Adelie1=0, Chinstrap1=1);

Y los pinguinos Gentoo como (Adelie1=0, Chinstrap1=0);

Para comprobarlo se va a repetir el apartado anterior utilizando solo esas dos variables auxiliares: