Vecinos más Cercanos KNN (k-Nearest Neighbor)

K-NN

- El principio detrás de los métodos del vecino más cercano es encontrar un número predefinido de muestras de entrenamiento más cercanas en distancia al nuevo punto y predecir la etiqueta a partir de ellas.
- El valor de **k** puede elegirse heurísticamente, generalmente no desea que sea tan alto que los votos se vuelvan ruidosos (en el extremo, si tiene n puntos de datos y establece k = n, simplemente elegirá la etiqueta más común en el conjunto de datos) , y desea elegirlo para que coincida con el número de clases (es decir, no comparten divisores comunes excepto 1)

Estimador de densidad

- Se espera encontrar la densidad de dispersión de los puntos.
- Se asume que los datos o la muestra de los datos es $X = \{x^t\}_{t=1}^{N}$.
- Se sabe que la función de distribución acumulativa $F(x)$ en el punto x es la proporción de puntos de muestra que son menores o iguales que x

$$F(x) = \frac{\#\{x^t \leq x\}}{N}$$

- $\#\{x^t \leq x\}$ es el número de instancias entrenamiento menores o iguales que x.
- Del mismo modo, el estimado para la función de densidad se puede calcular como $p(x) = \frac{1}{h}\left[\frac{\#\{x^t \leq x+h\} - \#\{x^t \leq x\}}{N}\right]$, donde h es el intervalo y las instancias x^t que caen en el intervalo son lo "suficientemente cercanas".

Estimador K-NN

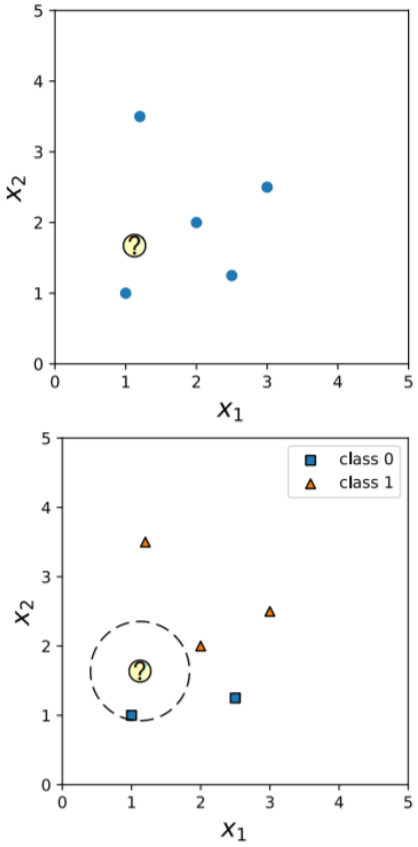
- k es el número de vecinos tomados en cuenta.
- Se define la distancia entre los puntos a y b como $|a-b|$, entonces para cada dato x: $d_1(x) \leq d_2(x) \leq \cdots \leq d_N(x)$.
- Como las distancias se ordenan ascendentemente, $d_1(x)$ es la muestra más cercana, $d_2(x)$ es la siguiente más cercana, y así sucesivamente.
- Como x^t son los puntos de los datos, se define $d_1(x) = min_t|x - x^t|$.
- Y si i es el índice de la muestra más cercana, e $i = \arg min_t|x - x^t|$, entonces $d_2(x) = min_{j \neq i}|x - x^t|$.
- Así el estimador de densidad K-NN es $\mathrm{p}(x) = \frac{K}{2Nd_k(x)}$

K-NN

- Para clasificar un dato, se usa el estimador de las densidades condicionales de clase $p(x|C_i)$.
- Para el estimador k-nn se tiene que $p(x|C_i) = \frac{k_i}{N_i V^k(x)}$.
- Done k_i es el número de vecinos más cercanos que pertenecen a C_i.
- Y $V^k(x)$ es el volumen en el espacio d-dimensional centrado en x, con el radio $r = \|x - x_{(k)}\|$. Donde $x_{(k)}$ es la k-esima observación más cercana a x (entre todos los vecinos de todas las clases de x). $V^k = r^d c_d$ con c_d como el volumen de la esfera unitaria en d-dimensión.
- x se asigna a la clase para la cual el discriminante toma su máximo.
- Entonces cada instancia de entrenamiento vota por su clase y no tiene efecto en otras clases.

Clasificación

- Entonces $p(C_i|x) = \frac{p(x|C_i)\mathrm{P}(C_i)}{p(x)} = \frac{k_i}{k}$
- Para un dato de entrada desconocido, el clasificador k-nn asigna el dato la clase que tiene la mayoría de los votos entre los k vecinos más cercanos.
- Todos los vecinos tienen igual voto, y se elige la clase que tiene el número máximo de votantes.
- Los empates se rompen arbitrariamente o se realiza una votación ponderada.
- Generalmente se considera que k es un número impar para minimizar la confusión es entre dos clases vecinas.



¡A codificar!