Auxiliar 6 - K-NN y Regresión

Cátedra: Inteligencia Artificial
Profesor: Gonzalo Rios
Auxiliar: Miguel Romero

30 de Mayo del 2011

1. Deseamos crear un sistema que detecte si un nuevo paciente tiene diabetes o no. Los datos de los pacientes disponibles son: Nombre, edad, nivel de glucosa (rango [0, 7]), tipo de sangre (entre T1, T2 y T3), indice A ([-100, 100]) e indice B ([-30, 30]). Queremos utilizar k-NN, para cierto k. Proponga un modelo de datos y una distancia para este sistema en los siguientes casos:
   (a) Dos pacientes cuyas edades no difiere en más de 5 años, parecen tener las mismas características frente a la diabetes.
   (b) Dos pacientes tendrán características similares frente a la diabetes si están en la misma época de sus vidas (1era, 2era, o 3era edad). Se tiene que el impacto de la época de la vida, nivel de glucosa, tipo de sangre, indice A e indice B, en la diabetes es un 15%, 40%, 25%, 5% y 15%, respectivamente.
2. Deseamos clasificar datos dentro de dos clases C_1 y C_2 (clasificación binaria). Cada dato es un punto en $\mathbb{R}^2$ y emplearemos la distancia euclideana.
   (a) Imagine que solo tenemos dos datos de entrenamiento $z_1 = (2, 3)$ y $z_2 = (5, 7)$, donde z_1 y z_2 están en la clase C_1 y C_2, respectivamente. Calcule explicitamente la frontera de decisión que entrega el algoritmo 1-NN.
   (b) Asuma que agregamos un nuevo dato $z_3 = (8, 6)$, el cual está en la clase C_2. Que sucede con la frontera de decisión de 1-NN? Explique graficamente.
3. Considere dos clases C_1 y C_2, datos de entrenamientos que son puntos en $\mathbb{R}^d$ y la distancia euclideanas. Sea j un numero natural par. La regla de clasificación j-th-NN es como sigue: Dada una nueva instancia a clasificar x, escogemos la clase C_1 si el j-esimo punto más cercano a x de la clase C_1, está mas cerca de x que el j-esimo punto mas cercano a x de la clase C_2. Escogemos C_2 en caso contrario.
   (a) Aplique el algoritmo 2-th-NN sobre la instancia $(3, 5)$ y los datos de entrenamiento $\{(3, 3, C_1), (1, 0, C_1), (-2, 4, C_1), (-1, 3, C_2), (5, 6, C_2), (4, 8, C_2)\}$.
   (b) Cual es la relación entre j-th-NN y k-NN, cuando $j = \frac{k+1}{2}$?.
4. Tenemos un conjunto de datos de entrenamiento $\{(\vec{x}_i, y_i) : i = 1, ..., m\}$, donde $\vec{x}_i \in \mathbb{R}^d$ e $y_i \in \mathbb{R}$.
   (a) Deduzca que el regresor lineal que minimiza el error cuadrático esta dado por $\hat{\beta} = (X^T X)^{-1} X^T y$, donde

$$X = \begin{bmatrix} 1 & x_1^1 & x_1^2 & \dots & x_1^d \\ 1 & x_2^1 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_m^1 & x_m^2 & \dots & x_m^d \end{bmatrix} \qquad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_m \end{pmatrix}$$

(b) Suponga que tiene dos atributos X_1 y X_2 de una persona, que toman valores en $\mathbb{R}$, y la variable Y que desea entender es la probabilidad de que la persona sea un asesino. Tenemos un conjunto de datos de entrenamiento $\{(x_i^1, x_i^2, y_i) : i = 1, .., m\}$. Cual es el problema con aplicar directamente regresión lineal? Como ocuparía la función $f(z) = \frac{1}{1+e^{-z}}$?