M2.2.2 Modelos Supervisados y No Supervisados

Programa Big Data y Business Intelligence

Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/

Deusto
Facultad de Ingeniería
Ingeniaritza Fakultatea

K-vecinos más cercanos (K nearest neighbors, KNN)

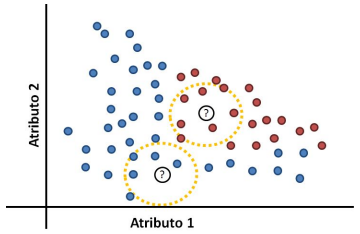
- Midiendo distancias entre los datos
- Estableciendo el valor de k
- Resolviendo empates

Introducción

- KNN representa el modelo por medio del conjunto de entrenamiento al completo
  - (Tan simple como eso)
- No hay más modelo que almacenar el conjunto de datos de entrenamiento
  - No es necesaria una etapa de aprendizaje
  - (Algunas implementaciones utilizan estructuras complejas para simplificar)

Introducción

- Principio “Dime con quién andas...”
  - Se basa en calcular la clasificación directamente a partir de los ejemplos
- Clasificar ejemplos a partir de los ejemplos más “cercanos”.
  - Necesitamos medir “Distancias” entre ejemplos.
  - En la mayoría de los casos, los ejemplos serán elementos numéricos
    - La distancia Euclídea puede ser una opción razonable.



Introducción

- Ejemplo de los métodos de aprendizaje “vago”
  - No hay modelo por detrás
- Cada dato de la muestra de entrenamiento se puede ver como un caso resuelto
- Para un dato nuevo, buscamos casos “más parecidos” y le aplicamos la misma solución

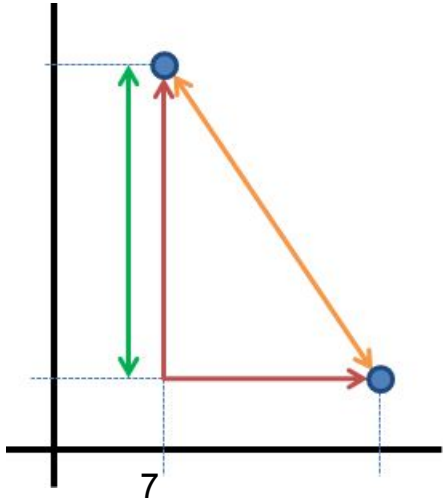
“if it walks like a duck, looks like a duck, and talks like a duck, it is probably a duck.

Midiendo Distancias

- Ciertos métodos de minería de datos se basan o necesitan calcular "distancias" entre diferentes datos
  - Y KNN es uno de ellos
    - Se basa en medir distancias entre los datos disponibles y el nuevo ejemplo
    - Para ello, mide la distancia entre el nuevo ejemplo y los ejemplos almacenados, y clasifica el nuevo por el voto de sus vecinos
  - Casos "más parecidos" = Casos a menor distancia
- ¿Os suena de algo?

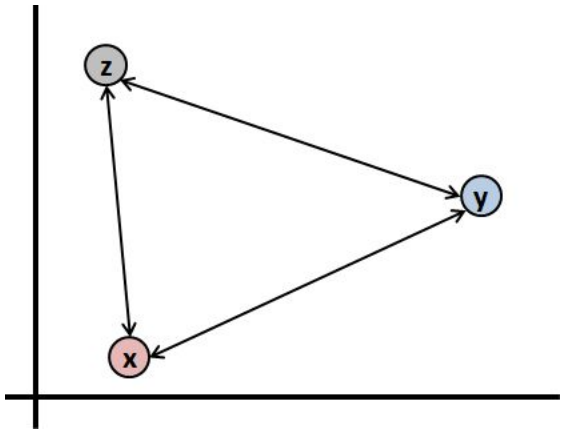
Midiendo Distancias

- Hay múltiples medidas de distancia
  - Las distancias más comunes son las euclídeas
    - Norma L2: raíz cuadrada de la suma de los cuadrados de las diferencias entre x e y en cada dimensión
      - La noción más común de distancia
    - Norma L1: suma de las diferencias en cada dimensión
      - Distancia de Manhattan: distancia si sólo te mueves a través de las coordenadas
    - Norma L∞: máxima distancia de las diferencias entre x e y en cada dimensión



Midiendo Distancias

- Una medida de distancia <d(x,y)> debe cumplir
  - Debe de ser mayor que cero
  - Es igual a cero si sólo si ambos elementos son iguales
  - Es simétrica (d(x,y)=d(y,x))
  - La distancia entre x e y es menor o igual que la distancia entre x y z más la distancia entre z e y
    - Desigualdad triangular
      - d(x,y)<d(x,z)+d(z,y)
      - d(x,z)<d(x,y)+d(y,z)
      - d(y,z)<d(y,x)+d(x,z)



Midiendo Distancias

- A quién se parece más Pedro?
  - (1) Pedro: Salario = 25.000, Edad= 30
  - (2) Juan: Salario = 27.000, Edad= 50
  - (3) Daniel: Salario = 20.000, Edad= 32
  - A Juan, que gana parecido pero es más senior.
  - A Daniel, que tiene una edad parecida pero cobra una cantidad menor
- Por qué?
- Cómo lo mido (numéricamente)?

Midiendo Distancias

- A quién se parece más Pedro?
  - (1) Pedro: Salario = 25.000, Edad= 30
  - (2) Juan: Salario = 27.000, Edad= 50
  - (3) Daniel: Salario = 20.000, Edad= 32
    - Parece lógico que Pedro se parece a Daniel
    - Pero si “pinto los datos y mido”
    - Pedro está más cerca de Juan
- Esto se debe a
  - 20 unidades (años) “no son nada” al lado de 2.000 unidades (euros)

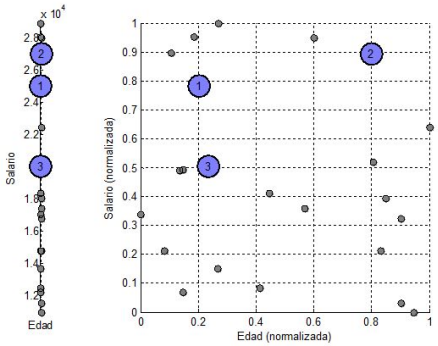
Midiendo Distancias

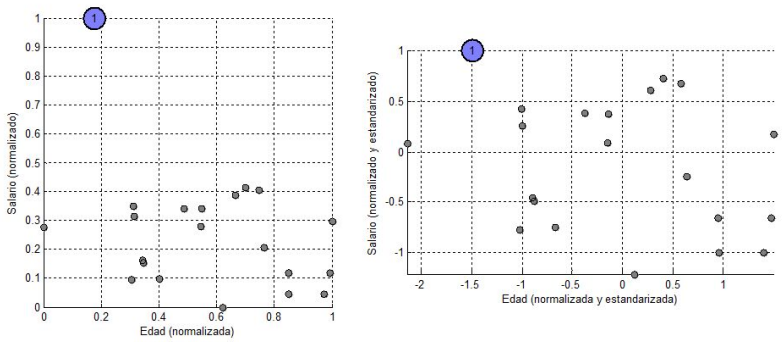
- Para medir la distancia entre las instancias de datos, es recomendable que todos los atributos estén en la misma escala
  - Normalización: escala los valores numéricos de manera que estén en el Rango [0,1]
  - Estandarización: hace que la distribución de los datos sea normal (En el sentido estadístico de la palabra)
    - Nota, hay otras muchas maneras de normalizar/estandarizar datos, estas dos son las (sumamente) más comúnes y utilizadas

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad x_{new} = \frac{x - \mu}{\sigma}$$

Midiendo Distancias

- Tanto normalizar como estandarizar tienen sus ventajas e inconvenientes:
  - Normalizar nos asegura que los valores estarán entre 0 y 1, para todos los atributos (Estandarizar, no)
  - Estandarizar mitiga el efecto negativo de los outliers, hace que los datos estén “mejor distribuidos”, lo que facilita la tarea de minería (Normalizar, no)





Midiendo Distancias

- Otras (de las muchas existentes) distancias
  - Hamming: número de elementos diferentes
    - Útil para datasets con atributos categóricos
  - Diferencia de ranking
    - Atributos ordinales
  - Distancia de editado
    - Cuando los datos son strings: El menor número de inserciones y borrados (y modificaciones, en ciertas ocasiones) necesarios para pasar de la cadena x a la cadena y
  - Matriz de distancias específica

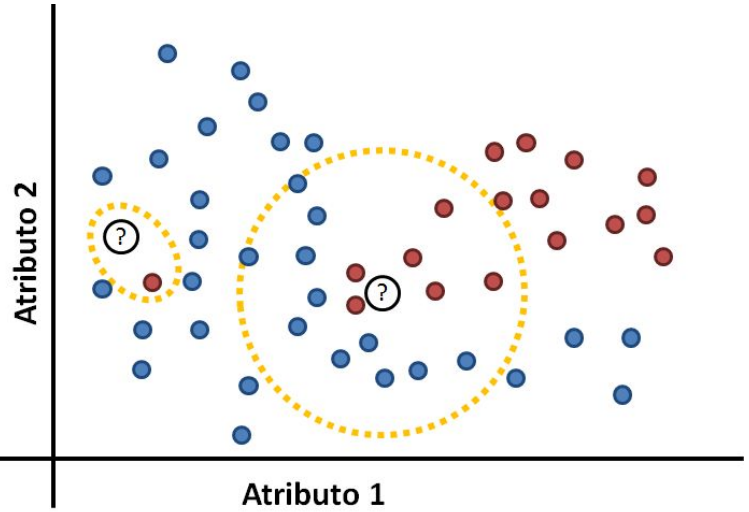
	Ingeniero	Empresariales	Abogado
Ingeniero	0	0.5	1
Empresariales	0.5	0	0.5
Abogado	1	0.5	0

Distancias en KNN

- Para ver los vecinos que son más parecidos, utilizamos medidas de distancia
  - Si los atributos no son numéricos:
    - Categorías → Dummy variables
    - Categorías (con orden) → numérico
    - Texto → Bag of words
  - Podemos usar medidas de distancia específicas
  - Si los atributos no tienen escala similar, podemos normalizar/estandarizar
- Otras cuestiones
  - ¿Qué valor de k usar?
  - ¿Cómo resolver empates?

Estableciendo el valor de K

- Si k es muy bajo
  - Puede ser muy sensible al ruido (y outliers)
    - Overfitting
- Si k es muy alto
  - Puede estar influenciado por la “mayoría no parecida”
    - ¿Y si K es igual al número de datos?
- Algunos criterios
  - K=raíz(N)
  - Probar diferentes K
    - Validación cruzada



Rompiendo empates

- ¿Qué hacer si entre los K vecinos más cercanos hay empate en las clases?
  - Intentar evitar el empate
    - Utilizar K impar si tenemos 2 clases
  - Si no se puede evitar (3 o más clases)
    - Tomar la clase del vecino más cercano
    - Tomar la clase con menor distancia promedio
    - Incrementar el valor de K (tomar un nuevo vecino) para romper el empate

Algunas variantes

- Variante en KNN:
  - Para cada clase, sumar la similitud (con el que se quiere clasificar) de cada ejemplo de esa clase que esté entre los k más cercanos. Devolver la clase con mayor puntuación.
    - Así un ejemplo cuenta más cuanto más cercano esté
- Adaptable a Regresión:
  - dar como salida el valor … de los k vecinos más cercanos
    - Medio
    - Mediano
    - Modal

Resúmen

- Ventajas:
  - (Muy) Simple
  - No hay que hacer presunciones sobre los datos
  - Entrenamiento rápido (o nulo)
- Desventajas
  - No produce un modelo.
  - Hay que elegir el número de vecinos a considerar
  - Tarda en clasificar
    - ¿Y si tenemos millones de datos?
  - La maldición de la dimensionalidad
    - Si tenemos muchos atributos, las distancias serán similares

Enrique Onieva
enrique.onieva@deusto.es
https://twitter.com/EnriqueOnieva
https://www.linkedin.com/in/enriqueonieva/