UNIDAD TEMÁTICA 4: Algoritmos No Lineales

Trabajo de Aplicación 8 – k-NN

Ejercicio 1

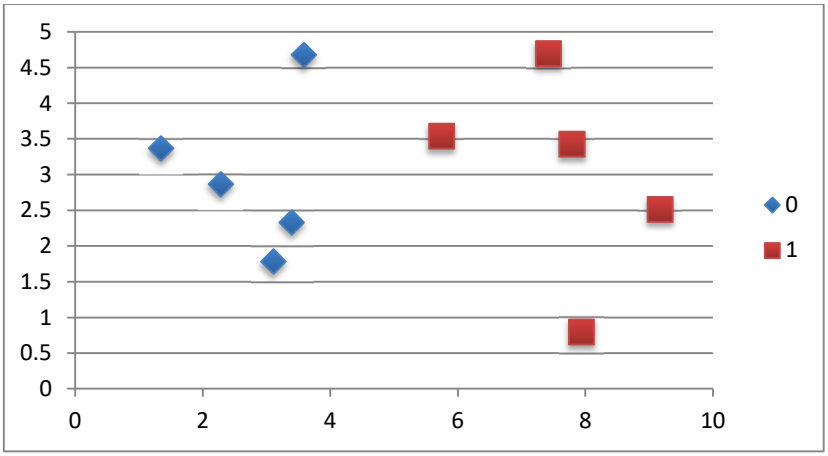
En este ejercicio practicaremos el algoritmo k-NN. Veremos cómo calcular la distancia euclidiana entre vectores de valores reales, y cómo utilizar esta distancia y el conjunto de entrenamiento para hacer predicciones sobre datos nuevos.

El Dataset

Este es un problema de clasificación binaria sencillo. El dataset tiene dos variables de entrada X1 y X2, y una variable de salida Y con dos clases, 0 y 1.

Abrir la planilla electrónica adjunta (TA8 -KVecinosMasCercanos.xlsx),

- graficar los datos de entrenamiento como puntos, con dos series,
  - para Y=0 e Y = 1
- ¿cómo se ven los ejemplos en cuanto a su clasificación?



Vemos que los ejemplos están bien separados en las clases (el ejemplo es artificial)

Distancia Euclidiana

KNN utiliza medidas de distancia para ubicar los ejemplos más cercanos. Una de las medidas utilizadas es la distancia euclidiana.

$$EuclideanDistance(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

Utlizando la planilla, calcularemos primero para el primer ejemplo.

DiferenciaCuadrada-1 = $(X1_1\text{-}X1_2)^2$ = 0.08034941698

DiferenciaCuadrada-2 = $(X2_1\text{-}X2_2)^2$ = 0.3022071889

Suma = DiferenciaCuadrada-1 + DiferenciaCuadrada-2 = 0.382556606

Distancia = RaizCuadrada (Suma) = 0.618511605

(observa que probablemente podríamos prescindir de esta última parte – la raíz cuadrada – pues es costosa computacionalmente)

Predicciones con el modelo KNN

Dada una nueva instancia o ejemplo para el cual queremos realizar una predicción, se eligen las k instancias con la menor distancia a la nueva instancia, para contribuir a la predicción.

Para tareas de clasificación, esto implica permitir que cada uno de los k miembros "vote" por la clase a que la nueva instancia pertenecería.

Aquí están los datos para la nueva instancia que queremos predecir:

X1 = 8.093607318 X2 = 3.365731514 Y = 1

Inclúyela en el gráfico realizado anteriormente, para visualizar dónde debe clasificar.

El primer paso es calcular la distancia euclidiana entre la nueva instancia de entrada y todas las instancias que tenemos en el dataset de entrenamiento.

Predicción Instancia	X1	X2	Y	(X1-X1)^2	(x2-X2)^2	Suma	Distancia
1	3.39353321	2.33127338	0	22.0906966	1.07010363	23.1608002	4.81256691
2	3.11007348	1.78153964	0	24.8356095	2.5096639	27.3452734	5.22927083
3	1.34380883	3.36836095	0	45.5597796	6.914E-06	45.5597865	6.749799
4	3.58229404	4.67917911	0	20.3519475	1.72514459	22.0770921	4.69862661

Asumiremos **k = 3**. Buscar los 3 ejemplos que presenten las menores distancias a la nueva instancia.

Ahora, hacer una predicción (en este caso, de clasificación). Podemos utilizar la función MODE (o MODA) para obtener el valor más frecuente (de las clases de los vecinos más cercanos)

Predicción = moda (clase (i))

1. ¿cuál será la predicción para este ejemplo? Registra los resultados

2. Crea un nuevo ejemplo que no sea tan evidente, a partir del gráfico de valores, y aplica el procedimiento para la predicción. Registra los resultados.
3. Repite el proceso, para ambos ejemplos, variando el valor de k
    a. Registra los resultados para k = 1 y k = 2

Ejercicio 2

Utilizaremos k-nn para clasificar las plantas de la especie “Iris”

Preparación de datos

1. Crea un nuevo proceso en RapidMiner
2. Descarga de UCI el dataset “Iris” e impórtalo.
    - Observa que tiene 4 atributos (reales) y una variable de salida, polinomial, que clasifica los tipos de plantas en 3 clases diferentes
3. Realiza un gráfico bidimensional, tomando como ejes “petal_length” y “petal_width”, y como “color column” la clase. Observa la distribución de los ejemplos.
    - ¿qué consideraciones puedes hacer a priori, en base a esta observación? Remite los comentarios a la tarea.
4. ¿qué tareas de acondicionamiento / preparación de los datos deben efectuarse?
    - Registra en un documento de texto y aplícalas al dataset
5. Agrega un operador “Split Data” para particionar el conjunto original en 2 subconjuntos del mismo tamaño, en forma aleatoria. Uno se usará para entrenamiento y el otro para test

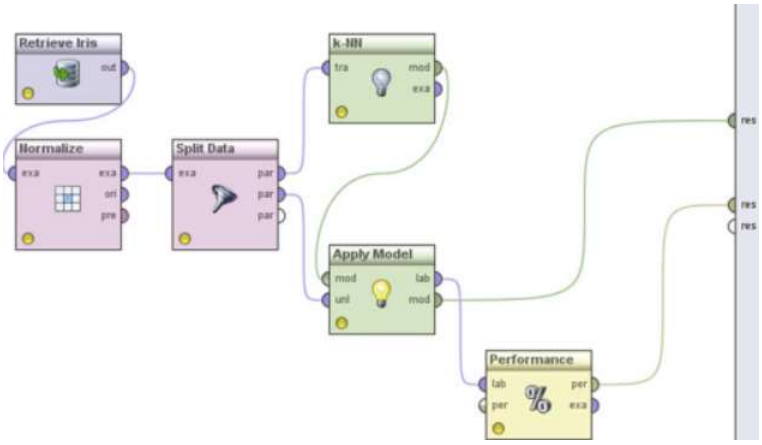
Operador de modelo y parámetros

El operador KNN de RapidMiner tiene algunos parámetros que se pueden configurar. Observa los mismos y resume en un documento de texto las principales características:

- k – tiene un valor por defecto de 1. Cámbialo a 3.
- Voto ponderado (Weighted Vote) - ¿cómo funciona? Toma nota de esto.
- Tipos de medición. RapidMiner tiene incluidas varias funciones para medición de distancia, que están agrupadas en Tipos de Medición
    - ¿Cuáles son estos tipos?
    - ¿qué características tiene cada uno?
- Funciones de medición. Observa las que están disponibles
    - Registra los nombres
    - ¿cómo funciona cada una?

Evaluación

Agrega un operador “Apply Model” y un “Performance (classification)”, y conecta los ports en forma apropiada para observar y comparar los resultados.



Ejecución e interpretación

1. Ejecutar el modelo y observar los resultados.
    a. Modelo k-nn: el modelo es simplemente todo el conjunto de entrenamiento.
    b. Vector de performance: matriz de confusión
2. Prueba con al menos 2 funciones de medición y valores de k diferentes. Realiza una matriz con estos datos, indicando los valores de exactitud de predicción alcanzados en cada caso.
3. En un POSTER, resume los hallazgos en función de los diferentes valores de k y de las funciones de distancia utilizadas. Explica estos resultados.

DISCUSION

Los equipos evaluarán los POSTERS de los demás equipos, y luego seguirá una discusión acerca de los mejores enfoques y resultados.