NOMBRE: Díaz Padilla Danny Sebastián
FECHA: 07-01-2020
PROFESOR: Maria Hallo.
TÍTULO: "Ejercicios de los capítulos 1,2,3,4,5,7 del libro de Principles of Datamining, Bramer, third edition".

Para el Capítulo 1 no hay ejercicios de autoevaluación

Ejercicios de autoevaluación para el Capítulo 2

1. **¿Cuál es la diferencia entre los datos etiquetados y no etiquetados?**

Los datos etiquetados tienen un atributo especialmente designado y el objetivo es usar los datos proporcionados para predecir el valor de ese atributo para instancias que todavía no han sido vistas.

Mientras que los datos no etiquetados no tienen un atributo especial designado y su objetivo es usar los datos para encontrar la mayor cantidad de relaciones entre sí.

2. **La siguiente información se encuentra en una base de datos de empleados.**

Nombre, fecha de nacimiento, sexo, peso, altura, estado civil, número de hijos

¿Cuál es el tipo de cada variable?

Nombre: Nominal

Fecha de nacimiento: ordinal

Sexo: binario

Peso: ratio-scaled

Altura: ratio-scaled

Estado civil: nominal

Número de hijos: entero

3. **Dé dos formas de lidiar con los valores de datos faltantes.**

- Uno, eliminando toda la instancia
- Dos, reemplazarlos con un promedio de la fila o el valor más frecuente

Ejercicios de autoevaluación para el Capítulo 3

1. Usando el algoritmo de clasificación Naıve Bayes con el conjunto de datos del tren, calcule

La clasificación más probable para las siguientes instancias invisibles.

weekday	summer	high	heavy	????

Probabilidad de que class = on time
0.70 × 0.64 × 0.43 × 0.29 × 0.07 = 0.0039
Probabilidad de que class = late
0.10 × 0.5 × 0 × 0.5 × 0.5 = 0
Probabilidad de que class = very late
0.15 × 1 × 0 × 0.33 × 0.67 = 0
Probabilidad de que class = cancelled
0.05 × 0 × 0 × 1 × 1 = 0
El valor más grande es 0.39% para 'on time'

sunday	summer	normal	slight	????

Probabilidad de que class = on time
0.70 × 0.07 × 0.43 × 0.36 × 0.57 = 0.0043
Probabilidad de que class = late
0.10 × 0 × 0 × 0.5 × 0 = 0
Probabilidad de que class = very late
0.15 × 0 × 0 × 0.67 × 0 = 0
Probabilidad de que class = cancelled
0.05 × 0 × 0 × 0 × 0 = 0
El valor más grande es 0.43% para la clase 'on time';

2. Usando el conjunto de entrenamiento que se muestra en la Figura 3.5 y la distancia euclidiana. Medir, calcular los 5 vecinos más cercanos de la instancia con primero y segundos atributos 9.1 y 11.0, respectivamente.

Para esto se utilizó la fórmula
RAIZ((9.1-A1)^2 + (11-A2)^2)
Donde esa letra representa a Atributo. Los 5 vecinos mas cercanos están pintados con rojo

Atributo 1	Atributo 2	Clase	Distancia respecto al objetivo
0,80	6,30	–	9,5383
1,40	8,10	–	8,2280
2,10	7,40	–	7,8715
2,60	14,30	+	7,2897
6,80	12,60	–	2,8018
8,80	9,80	+	1,2369
9,20	11,60	–	0,6083
10,80	9,60	+	2,2023
11,80	9,90	+	2,9155
12,40	6,50	+	5,5803
12,80	1,10	–	10,5688

14,00	19,90	–	10,1597
14,20	18,50	–	9,0697
15,60	17,40	–	9,1220
15,80	12,20	–	6,8066
16,60	6,70	+	8,6452
17,40	4,50	+	10,5423
18,20	6,90	+	9,9810
19,00	3,40	–	12,4808
19,60	11,10	+	10,5005

Ejercicios de autoevaluación para el Capítulo 4

1. ¿Cuál es la condición de adecuación en las instancias de un conjunto de capacitación?

No pueden pertenecer dos instancias con los mismos valores de todos los atributos diferentes clases.

2. ¿Cuáles son las razones más probables para que la condición no se cumpla para un determinado conjunto de datos?

El ruido
Los valores faltantes.

3. ¿Cuál es el significado de la condición de adecuación para la generación automática de reglas utilizando el algoritmo TDIDT?

Siempre que se cumpla la condición de adecuación, se garantiza el algoritmo TDIDT para dar un árbol de decisión correspondiente al conjunto de entrenamiento.

4. ¿Qué sucede si el algoritmo TDIDT básico se aplica a un conjunto de datos para que la condición de adecuación no se aplica?

Se alcanzará una situación en la que se ha generado una rama a la máxima longitud posible, con un término para cada uno de los atributos, pero el correspondiente subconjunto del conjunto de entrenamiento todavía tiene más de una clasificación.

Ejercicios de autoevaluación para el Capítulo 5

1. Al construir una hoja de cálculo o de otra manera, calcule lo siguiente para conjunto de datos de grados dado en la Sección 4.1.3, Figura 4.3:

a) la entropía inicial Estart

Estart = - (6/26) log2 (6/26) - (20/26) log2 (20/26) = 0.7793.

b) la entropía promedio ponderada Nueva de los conjuntos de entrenamiento (sub) resultantes de dividirse en cada uno de los atributos SoftEng, Arin, HCI, CSA y Proyecto a su vez y el valor correspondiente de Ganancia de información en cada caso. Con estos resultados, verifique que el atributo que elegirá el Algoritmo TDIDT para la primera división en los datos usando la selección de entropía el criterio es SoftEng.

Partiendo en SoftEng
SoftEng = A
Proporciones de cada clase: PRIMERO 6/14, SEGUNDO 8/14
Entropía = - (6/14) log2 (6/14) - (8/14) log2 (8/14) = 0.9852

SoftEng = B
Proporciones de cada clase: PRIMERO 0/12, SEGUNDO 12/12
Entropía = 0 [todas las instancias tienen la misma clasificación]
Entropía promedio ponderada Enew = (14/26) × 0.9852 + (12/26) × 0 = 0.5305
Ganancia de información = 0.7793 - 0.5305 = 0.2488

División en ARIN

ARIN = A
Proporciones de cada clase: PRIMEROS 4/12, SEGUNDO 8/12
Entropía = 0.9183

ARIN = B
Proporciones de cada clase: PRIMERO 2/14, SEGUNDO 12/14
Entropía = 0.5917
Entropía promedio ponderada Enew = (12/26) × 0.9183+ 14/26 × 0.5917 = 0.7424
Ganancia de información = 0.7793 - 0.7424 = 0.0369

División en HCl

HCl = A
Proporciones de cada clase: PRIMERO 1/9, SEGUNDO 8/9
Entropía = 0.5033

HCl = B
Proporciones de cada clase: PRIMERO 5/17, SEGUNDO 12/17
Entropía = 0.8740
Entropía promedio ponderada Enew = (9/26) × 0.5033 + (17/26) × 0.8740 = 0.7457
Ganancia de información = 0.7793 - 0.7457 = 0.0337

División en CSA

CSA = A
Proporciones de cada clase: PRIMERO 3/7, SEGUNDO 4/7
Entropía = 0.9852

CSA = B
Proporciones de cada clase: PRIMERO 3/19, SEGUNDO 16/19
Entropía = 0.6292
Entropía promedio ponderada Enew = (7/26) × 0.9852 + (19/26) × 0.6292 = 0.7251
Ganancia de información = 0.7793 - 0.7251 = 0.0543

División en proyecto

Proyecto = A
Proporciones de cada clase: PRIMERO 5/9, SEGUNDO 4/9
Entropía = 0.9911

Proyecto = B
Proporciones de cada clase: PRIMERO 1/17, SEGUNDO 16/17
Entropía = 0.3228
Entropía promedio ponderada Enew = (9/26) × 0.9911 + (17/26) × 0.3228 = 0.5541
Ganancia de información = 0.7793 - 0.5541 = 0.2253

El valor máximo de ganancia de información es para el atributo SoftEng.

2. Sugerir razones por las cuales la entropía (o ganancia de información) es una de las más efectivas métodos de selección de atributos cuando se usa la generación de árbol TDIDT algoritmo.

El algoritmo TDIDT conduce a un árbol de decisión donde todos los nodos tienen entropía cero. Reducir la entropía promedio tanto como sea posible en cada paso parecería una manera eficiente de lograr esto en un número relativamente pequeño de pasos El uso de minimización de entropía (o maximización de ganancia de información) generalmente parece conducir a un pequeño árbol de decisión en comparación con otros Criterios de selección de atributos. El principio de la Navaja de Occam sugiere que Es muy probable que los árboles sean los mejores, es decir, que tengan el mayor poder predictivo.

Ejercicios de autoevaluación para el Capítulo 7

1. Calcule la precisión predictiva y el error estándar correspondiente a la matrices de confusión dadas en las Figuras 7.14 y 7.15. Para cada conjunto de datos, estado el rango en el que se puede esperar el verdadero valor de la precisión predictiva mentir con probabilidad 0.9, 0.95 y 0.99.

Para figura 7.14

El número de predicciones correctas es 127 y el número total de instancias es 135.
Tenemos p = 127/135 = 0.9407, N = 135, entonces el error estándar es p × (1 - p) / N = 0.9407 × 0.0593 / 135 = 0.0203.
Se puede esperar que el valor de la precisión predictiva se encuentre en los siguientes rangos:
probabilidad 0.90: de 0.9407 - 1.64 × 0.0203 a 0.9407 + 1.64 × 0.0203, es decir de 0,9074 a 0,9741
probabilidad 0.95: de 0.9407 - 1.96 × 0.0203 a 0.9407 + 1.96 × 0.0203, es decir de 0.9009 a 0.9806
probabilidad 0.99: de 0.9407 - 2.58 × 0.0203 a 0.9407 + 2.58 × 0.0203, es decir de 0.8883 a 0.9932

Para Figura 7.15

El número de predicciones correctas es 149 y el número total de instancias es 214.
Tenemos p = 149/214 = 0.6963, N = 214, entonces el error estándar es p × (1 - p) / N = 0.6963 × 0.3037 / 214 = 0.0314.
Se puede esperar que el valor de la precisión predictiva se encuentre en los siguientes rangos:
probabilidad 0.90: de 0.6963 - 1.64 × 0.0314 a 0.6963 + 1.64 × 0.0314, es decir de 0.6447 a 0.7478
probabilidad 0.95: de 0.6963 - 1.96 × 0.0314 a 0.6963 + 1.96 × 0.0314, es decir de 0.6346 a 0.7579
probabilidad 0.99: de 0.6963 - 2.58 × 0.0314 a 0.6963 + 2.58 × 0.0314, es decir de 0.6152 a 0.7774

2. Sugerir algunas tareas de clasificación para las que sea falso positivo o falso negativo Las clasificaciones (o ambas) serían indeseables. Para estas tareas, qué proporción de falsas negativas (positivas) clasificaciones estarías dispuesto aceptar para reducir la proporción de falsos positivos (negativos) a ¿cero?

Las clasificaciones falsas positivas llegan a ser muy costosas en aplicaciones como

- la predicción de equipos que fallarán en el futuro cercano, por el mantenimiento.
- Clasificaciones falsas de individuos como probables criminales o terroristas
- Exámenes médicos
- Desastres naturales

La aceptación de falsos depende de la persona y del estudio, para medicina no podría ser mayor a 5% y en otros casos depende del problema