UNIVERSIDAD AUSTRAL | INGENIERÍA

Introducción a Data Mining

Trabajo Práctico Nº 3 – Clasificación Basada en Reglas – Métodos Bayesianos

Nota: Es importante que realice estos ejercicios sin emplear software de mineria de datos

1. Considerar un problema de clasificación binaria con el siguiente conjunto de atributos y correspondientes valores:
   - Acondicionador de aire = {funciona, descompuesto}
   - Motor = {bien, mal}
   - Kilometraje = {alto, medio, bajo}
   - Oxido = {si, no}

   Suponga que un clasificador basado en reglas produce el siguiente conjunto de

Reglas:

Kilometraje = alto → Valor = bajo
Kilometraje = bajo → Valor = alto
Oxido=si → Valor = bajo
Oxido=no, Motor = bien → Valor = alto
Oxido=no, Motor = mal→ Valor = bajo
Acondicionador de aire = funciona, Motor = bien → Valor = alto
Acondicionador de aire = funciona, Motor = mal → Valor = bajo
Acondicionador de aire = descompuesto → Valor = bajo

a) Explique si las reglas son mutuamente exclusivas.

b) Explique si el conjunto de reglas es exhaustivo

c) ¿Se requiere un orden para este conjunto de reglas?

d) ¿Se necesita una clase default para el conjunto?

2. El algoritmo RIPPER es una extensión de un algoritmo anterior llamado IREP. Ambos algoritmos aplican el método de poda de error reducido para determinar si una regla necesita ser podada. El método de poda de error reducido usa un conjunto de validación para estimar el error de generalización de un clasificador. Considere el siguiente par de reglas:

$$R_1: A \rightarrow C$$
$$R_2: A \wedge B \rightarrow C$$

R_2 se obtiene agregando un nuevo conjunto B al lado derecho de R_1.

Se deberá determinar si R_2 es preferible a R_1, desde la perspectiva del crecimiento de reglas y la poda de reglas. Para determinar si una regla debería ser podada, IREP computa la siguiente medida:

$$V_{IREP} = (p + (N - n)) / (P + N)$$

Donde P es el número total de ejemplos positivos en el conjunto de validación, N es el número total de ejemplos negativos en el mismo conjunto, p es el número de ejemplos positivos en el conjunto de validación cubierto por la regla y n es el número de ejemplos negativos en el conjunto de validación cubierto por la regla. V_{IREP} es similar a la precisión de validación para el conjunto de validación. IREP favorece las reglas que tienen valores más altos de V_{IREP}. Por otra parte, RIPPER aplica la siguiente medida para determinar si una regla debería ser podada:

$$V_{RIPPER} = (p - n) / (p + n)$$

a) Suponga que R_1 es cubierta por 350 ejemplos positivos y 150 ejemplos negativos, mientras que R_2 es cubierta por 300 ejemplos positivos y 50 negativos. Calcule la ganancia de información de FOIL para la regla R_2 con respecto a R_1.

b) Considere un conjunto de validación que contiene 500 ejemplos positivos y 500 negativos. Para R_1, suponga que el numero de ejemplos positivos cubiertos por la regla es 200 y el número de ejemplos negativos cubiertos es 50. Para R_2, suponga que los ejemplos positivos cubiertos son 100 y el número de ejemplos negativos cubiertos es 5. Calcule V_{IREP} para ambas reglas. ¿Cuáles reglas prefiere IREP?

c) Calcule V_{RIPPER} para el problema previo. ¿Cuáles reglas prefiere RIPPER?

3. C4.5rules es una implementación de un método indirecto para generar reglas desde un árbol de decisión. RIPPER es una implementación de un método directo para generar reglas directamente desde los datos.

a) Discuta las fortalezas y debilidades de ambos métodos.

b) Considere un dataset que tiene una gran diferencia en el tamaño de clase. ¿Cuál método (entre C4.5rules y RIPPER) es mejor en términos de hallar reglas de mayor precisión para las clases pequeñas?

4. Considere un training set que contiene 100 ejemplos positivos y 400 negativos. Para cada una de las siguientes reglas candidatas,

R_1: A $\rightarrow$ + (cubre 6 ejemplos positivos y 1 negativo)
R_2: B $\rightarrow$ + (cubre 40 ejemplos positivos y 10 negativos)
R_3: C $\rightarrow$ + (cubre 100 ejemplos positivos y 90 negativos)

Determine la mejor y la peor regla candidata de acuerdo con:

a) Precisión de la regla (accuracy)

b) Ganancia de información de FOIL.

c) La ratio de verosimilitud.

d) La medida de Laplace.

e) La medida de m-estimación (con k = 2 y $p_+ = 0.2$).

5. Suponga que la fracción de estudiantes no graduados que fuman es 15% y la fracción de estudiantes graduados que fuman es 23%. Si 1/5 de los estudiantes son graduados y el resto son no graduados, ¿Cuál es la probabilidad que un estudiante que fuma sea graduado?

6. Dado el dataset siguiente:

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

a) Estimar las probabilidades condicionales para P(A|+), P(B|+), P(C|+), P(A|-), P(B|-), P(C|-)

b) Usar las probabilidades condicionales de a) para predecir el rotulo de clase para una nueva instancia definida por A = 0, B = 1, C= 0 usando Naive Bayes.

c) Estimar las probabilidades condicionales usando m-estimate con p = ½ y m = 4.

d) Repetir b) usando las probabilidades condicionales halladas en c).

e) Comparar los dos métodos.