Aprendizaje Automático

Práctica metodología experimental: creación y evaluación de hipótesis (y poda de árboles)

Contenido

- Conjuntos de datos
- Algoritmos
- Hold out
- Validación cruzada de 10 particiones
- Elaboracion de tablas comparativas y discusión de resultados
- Contenido de la memoria

Conjuntos de datos

- Soybean
  - https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)
  - 683 instancias
  - 36 atributos (35 + clase)
  - 19 clases
- Vote
  - https://archive.ics.uci.edu/ml/datasets/congressional+voting+records
  - 435 instancias
  - 17 atributos (16 + clase)
  - 2 clases
- Thoracic_surgery
  - http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data
  - 470 instancias
  - 17 atributos (1+ clase)
  - 2 clases
- Realizar todos los experimentos sobre los tres conjuntos de datos

Algoritmos

- Árboles : J48, opciones por defecto, implementación de C4.5 en Weka
- Árboles sin podar: añadir opciones *collapseTree: false; subtreeRaising: False; unpruned: True (asimilable a ID3 con discretización de atributos continuos)*

- Para cada conjunto de datos, y algoritmo, entrenar y evaluar la tasa de error con los métodos que se piden

- En los métodos con repetición, sobre distintas particiones.
- Para ello, modificar la semilla para la generación de números aleatorios.
  - En Weka Explorer, Pestaña *Classify,* botón *More Options…*

Ejercicio Previo

- Para cada conjunto de datos, obtener una muestra aleatoria con 50 instancias para entrenar; utilizar las restantes muestras para estimar la tasa de error.
- Aplicar cada algoritmo sobre los conjuntos de datos así generados

Datos	**Algoritmo**	**Método: 50 T, resto**		
		Tasa error	**Desviación estándar**	**Intervalos**
Soybean_50	J48			
	Sin podar			
Vote_50	J48			
	Sin podar			
Thoracic_surgery_50	J48			
	Sin podar			

Hold out

- Realizar un experimento de Hold out 2/3 - 1/3, calculando la tasa de error, la desviación estándar y el intervalo de confianza del 95%
  - Asumir que hay suficientes datos y aproximar distribución binomial por distribución normal para calcular los intervalos de confianza
  - Utilizar la semilla aleatoria por defecto (valor 1).

Datos	Algoritmo	Método: Hold out		
		Tasa error	Desviación estándar	Intervalos
Soybean	J48			
	Sin podar			
Vote	J48			
	Sin podar			
Thoracic_ surgery	J48			
	Sin podar			

Hold out repetido (I)

- Realizar tres experimentos adicionales de Hold out 2/3 - 1/3, anotando la tasa de error de cada experimento
  - Utilizar tres semillas aleatorias diferentes en cada experimento (y distintas del valor por defecto)

Datos	Algoritmo	Tasa de error		
		2	3	4
Soybean	J48			
	Sin podar			
Vote	J48			
	Sin podar			
Thoracic_ surgery	J48			
	Sin podar			

Hold out repetido (II)

- Con todos los experimentos de hold out repetido determinar la tasa de error, la varianza y el intervalo de confianza para cada conjunto de datos y algoritmo

Datos	Algoritmo	Método: Hold out repetido		
		Tasa error	Desviación estándar	Intervalos
Soybean	J48			
	Sin podar			
Vote	J48			
	Sin podar			
Thoracic_ surgery	J48			
	Sin podar			

Validación cruzada 10 particiones

- Se proporcionan los resultados de los experimentos de validación cruzada para los tres conjuntos de datos
- Tasa de acierto sobre cada partición
- Semilla aleatoria por defecto (1)
- Algoritmo (1): J48; Algoritmo (2): J48 sin podar
- Determinar la tasa de error, la varianza y el intervalo de confianza para cada conjunto de datos y algoritmo

Datos	**Algoritmo**	**Método: 10 XV**		
		Tasa error	**Desviación estándar**	**Intervalos**
Soybean	J48			
	Sin podar			
Vote	J48			
	Sin podar			
Thoracic_ surgery	J48			
	JRIP			

Validación cruzada 10 particiones Soybean

- Tasa de acierto para Soybean, sobre cada partición

Dataset	(1) trees.J4	(2) trees
1	(10) 91.16	90.58
2	(10) 93.33	92.03
3	(10) 91.74	89.86
4	(10) 91.47	90.15
5	(10) 90.44	89.26
6	(10) 91.62	90.88
7	(10) 92.06	91.03
8	(10) 93.68	91.91 *
9	(10) 90.44	90.15
10	(10) 91.91	91.91

Validación cruzada 10 particiones Vote

- Tasa de acierto para Vote, sobre cada partición

Dataset	(1) trees.J4	(2) trees
1	(10) 97.05	96.36
2	(10) 96.14	95.45
3	(10) 97.05	95.91
4	(10) 97.05	96.59
5	(10) 97.05	97.05
6	(10) 95.81	94.88
7	(10) 96.28	95.58
8	(10) 96.74	96.05
9	(10) 95.58	93.95
10	(10) 96.98	95.81

Validación cruzada 10 particiones Thoracic_surgery

- Tasa de acierto para Thoracic_surgery, sobre cada partición

Dataset	(1) trees.J4	(2) trees
1	(1) 82.98	80.85 *
2	(1) 85.11	72.34 *
3	(1) 85.11	78.72 *
4	(1) 85.11	80.85 *
5	(1) 85.11	74.47 *
6	(1) 82.98	80.85 *
7	(1) 85.11	72.34 *
8	(1) 85.11	74.47 *
9	(1) 82.98	78.72 *
10	(1) 85.11	76.60 *

Validación cruzada repetida

- Realizar tres experimentos de validación cruzada de 10 particiones, anotando el error medio obtenido
  - Utilizar tres semillas aleatorias diferentes en cada experimento (y distinta del valor por defecto)

Datos	Algoritmo	Tasa de error		
		2	3	4
Soybean	J48			
	Sin podar			
Vote	J48			
	Sin podar			
Thoracic_ surgery	J48			
	Sin podar			

Tablas comparativas de la estimación del error

- Para cada conjunto de datos, elaborar una tabla con la tasa de error y la desviación estándar (si se ha estimado) estimada con cada método.

Algoritmo	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48				
Sin podar				

- Discutir los resultados
  - Para cada par datos-algoritmo examinar la variación de la tasa de error y la desviación estándar según el método empleado
  - Para cada conjunto de datos, qué método induce clasificadores con menor tasa de error
  - Para cada conjunto de datos, los tamaños de los árboles podados y sin podar

Contenido de la memoria

1. Descripción de los conjuntos de datos
2. Descripción de los algoritmos (no hay que describir C4.5, es un estándar; no hay que describir Weka, pero sí indicar las herramientas que utilizáis)
3. Experimentos con las muestras de 50 instancias de cada conjunto de datos (ejercicio previo)
4. Experimentos de hold out sin repetición
5. Experimentos de hold out con repetición
6. Experimentos de validación cruzada sin repetición
7. Experimentos de validación cruzada con repetición
8. Tablas comparativas y discusión de resultados
9. Referencias

(Y por supuesto, una carátula con el título de la práctica y vuestros nombres y apellidos)