Aprendizaje Automático

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

Contenido

- Conjuntos de datos
- Algoritmos
- Ejercicio inicial
- Hold out
- Hold oud repetido
- Validación cruzada de 10 particiones
- Validación cruzada repetida
- Elaboración de tablas comparativas y discusión de resultados
- Descripción del conjunto de datos
- Preguntas sobre validación cruzada repetida
- 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
- Realizar todos los experimentos sobre los dos 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 inicial

- 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			

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_50	J48			
	Sin podar			
Vote_50	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			

Hold out repetido (II)

- Con los 4 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			

Validación cruzada 10 particiones

- Se proporcionan los resultados de los experimentos de validación cruzada para los dos 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			

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 repetida (I)

- 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			

Validación cruzada repetida (II)

- Con los 4 experimentos de validación cruzada repetida determinar la tasa de error, la varianza y el intervalo de confianza para cada conjunto de datos y algoritmo
- **Atención**: **asumir como experimento base cada proceso de validación cruzada**

Datos	**Algoritmo**	**Método: Validación cruzada repetida**		
		Tasa error	**Desviación estándar**	**Intervalos**
Soybean	J48			
	Sin podar			
Vote	J48			
	Sin podar			

Tablas comparativas de la estimación del error

- Para cada conjunto de datos, elaborar una tabla con la tasa de error, la desviación estándar y los intervalos estimados con cada método.

Algoritmo	50 instan. entrenam.	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48					
Error					
Desviación					
Intervalos					
Sin podar					
Error					
Desviación					
Intervalos					

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 algoritmo induce clasificadores con menor tasa de error
- Para cada conjunto de datos, los tamaños de los árboles podados y sin podar (profundidad, número de nodos)
  - Solo se inducen dos árboles diferente para cada algoritmo y conjunto de datos: el que se crea con el conjunto de 50 instancias y el que se crea con cualquiera de los otros métodos de estimación de al tasa de error

Descripción del conjunto de datos

- En esta asignatura os vamos a pedir la información mínima necesaria para describir el conjunto de datos:
  - Nombre y origen del conjunto de datos (con enlace a la página de descarga, si disponible)
  - Número de instancias totales
  - Numero de atributos, indicando cuántos son numéricos y cuántos nomínales o categóricos. (También si hay atributos booleanos)
  - **Descripción de la clase.** Etiquetas de clase y su distribución

- Esta información se puede completar con una descripción mas detallada de cada atributo:
  - Rango, valores ausentes, valores inválidos, máximo, mínimo, media mediana, varianza y distribución para los atributos continuos
  - Rango, valores ausentes, valores inválidos, moda y distribución para los atributos nominales

Preguntas sobre validación cruzada repetida

- En esta práctica, en la validación cruzada repetida, hemos considerado como experimento base cada proceso de validación cruzada (Método 1)
- Sin embargo, es más habitual considerar como experimento base cada proceso de entrenamiento y validación sobre cada capa (*fold)*. (Método 2)

- Contestar razonadamente las siguientes preguntas
  - ¿Qué tasa de error se obtendría con el método 2?
  - ¿Cómo espera que varíe la estimación de la varianza con el método 2 frente al método 1?
  - ¿Y los intervalos de confianza?

- Si tiene curiosidad, puede comprobarlo con Weka Experimenter

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. Preguntas sobre validación cruzada
10. Referencias

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