Universidad de Valladolid

Práctica 4 - TAA

Pablo Martín de Benito

March 28, 2024

Contents

1 Método 50T, Resto **2**

2 Método Hold Out 2/3 - 1/3 **2**

3 Método Hold Out Repetido **3**

4 Validación Cruzada 10 Particiones **4**

5 Validación Cruzada Repetida **4**

6 Comparativas de la estimación del error **5**
6.1 Conjunto de datos - soybean 5
6.2 Conjunto de datos - vote 5

7 Conjuntos de datos **6**

8 Preguntas sobre validación cruzada **6**
8.1 ¿Qué tasa de error se obtendría con el método 2? 6
8.2 ¿Cómo espera que varíe la estimación de la varianza con el método 2 frente al método 1? 6

1 Método 50T, Resto

Datos	Algoritmo	Método 50T, resto			
		Tasa Error	Desviación Estándar	Intervalo	
soybean	j48	0,411671924290221	0,0195606623587688	0,373333026067034	0,450010822513408
	sin podar	0,416403785488959	0,0195934860745905	0,378000552782762	0,454807018195156
Vote	j48	0,077720207253886	0,0136448245874235	0,050976351062536	0,104464063445236
	sin podar	0,077720207253886	0,0136448245874235	0,050976351062536	0,104464063445236

Para ello, hemos utilizado los siguientes porcentajes de entrenamiento y clasificación:

- 7,32064421669107 % para soybean
- 11,4942528735632 % para vote

Además hemos utilizado las fórmulas de desviación típica e intervalo normales.

- Desviación Estándar: $S_{e(h)} = (\frac{e_{S(h)}(1-e_{S(h)}}{n})^{1/2}$
- Intervalo de Confianza Normal: $[e_{S(h)} \pm Z_{1-\alpha} \quad S_{e(h)}]$

2 Método Hold Out 2/3 - 1/3

Datos	Algoritmo	Hold Out			
		Tasa Error	Desviación Estándar	Intervalo	
soybean	j48	0,103004291845494	0,012081503606538	0,0793245447766795	0,126684038914308
	sin podar	0,107296137339056	0,0123010984913333	0,0831859842960427	0,131406290382069
Vote	j48	0,0337837837837838	0,00920790734958442	0,0157362853785983	0,0518312821889693
	sin podar	0,0337837837837838	0,00920790734958442	0,0157362853785983	0,0518312821889693

Utilizando las siguientes fórmulas para rellenar la tabla.

- Desviación Estándar: $S_{e(h)} = (\frac{e_{S(h)}(1-e_{S(h)}}{n})^{1/2}$
- Intervalo de Confianza Normal: $[e_{S(h)} \pm Z_{1-\alpha} \quad S_{e(h)}]$

3 Método Hold Out Repetido

Realizamos tres experimentos más de Hold Out 2/3-1/3 y anotando la tasa de error en cada experimento. Utilizando tres semillas diferentes.

Porcentaje			
soybean	2	3	4
j48	9,78723404255319	11,965811965812	11,6883116883117
sin podar	18,93617021276596	11,965811965812	13,8528138528139
vote	2	3	4
j48	6,75675675675676	5,40540540540541	6,12244897959184
sin podar	5,40540540540541	5,40540540540541	6,12244897959184

Ahora, con los cuatro experimentos, determinamos las tasas de error, desviación típica y los intervalos.

Datos	**Algoritmo**	**Hold Out Repetido**			
		Tasa Error	**Desviación Estándar**	**Intervalo**	
soybean	j48	0,109354467203066	0,0105673992752356	0,0925393784261278	0,126169555980004
	sin podar	0,113711024413243	0,0206986099918583	0,0807749201261683	0,146647128700318
Vote	j48	0,0541574738003309	0,0146614150119994	0,0308278930206922	0,0774870545799696
	sin podar	0,0507790954219526	0,0118236710231369	0,0319649981455036	0,0695931926984016

Para obtener los resultados de la tabla, utilizamos las siguientes fórmulas.

- Tasa Error: $e(h) = [\sum_{i=1,k} e_i(h)]/k$
- Desviación Estándar: $S_{e(h)} = \sqrt{\frac{1}{k-1} \quad \sum_{i=1,k} (e_i(h) - e(h))}$
- Intervalo de Confianza T-Student: $[e(h) \pm t_{N,k-1} \quad S_{e(h)}/\sqrt{k}]$
- Siendo qt(0.975,3) = 3,182446 calculado con R.

4 Validación Cruzada 10 Particiones

Se nos proporcionan los resultados de los experimentos de validación cruzada para los dos conjuntos

Datos	Algoritmo	Validación Cruzada 10 Particiones			
		Tasa Error	Desviación Estándar	Intervalo	
soybean	j48	0,0821500000000001	0,0106491783720623	0,0745320372568655	0,0897679627431346
	sin podar	0,09224	0,00953312354081513	0,0854204147303841	0,0990595852696159
Vote	j48	0,0342699999999999	0,0057205380477325	0,0301777738961873	0,0383622261038125
	sin podar	0,0423699999999999	0,00881980725412975	0,0360606904254833	0,0486793095745165

Para obtener los resultados de la tabla, utilizamos las siguientes fórmulas.

- Tasa Error: $e(h) = [\sum_{i=1,k} e_i(h)]/k$
- Desviación Estándar: $S_{e(h)} = \sqrt{\frac{1}{k-1} \quad \sum_{i=1,k} (e_i(h) - e(h))^2}$
- Intervalo de Confianza T-Student: $[e(h) \pm t_{N,k-1} \quad S_{e(h)}/\sqrt{k}]$
- Siendo qt(0.975,9) = 2,262157 calculado con R.

Siendo K = 10 Particiones

5 Validación Cruzada Repetida

Realizamos tres experimentos más de validación cruzada, anotando el error medio obtenido.
Utilizando tres semillas diferentes para cada experimento.

Porcentaje			
soybean	2	3	4
j48	9,78723404255319	11,965811965812	11,6883116883117
sin podar	18,93617021276596	11,965811965812	13,8528138528139
vote	2	3	4
j48	6,75675675675676	5,40540540540541	6,12244897959184
sin podar	5,40540540540541	5,40540540540541	6,12244897959184

Con los cuatro experimentos de validación cruzada repetida determinamos la tasa de error, la desviación estándar y el intervalo de confianza para cada conjunto de datos y algoritmo.

Datos	Algoritmo	Validación Cruzada Repetida			
		Tasa Error	Desviación Estándar	Intervalo	
soybean	j48	0,0882352941176471	0,033851824720493	0,0791987745400279	0,0972718136952663
	sin podar	0,0948316283034953	0,0350435294569074	0,0854769910684007	0,10418626553859
Vote	j48	0,0351083509513742	0,0269968624124527	0,0279017179932183	0,0423149839095301
	sin podar	0,0403012684989429	0,0315414221939052	0,0318814955172923	0,0487210414805935

Para obtener los resultados de la tabla, utilizamos la siguientes fórmulas.

- Tasa Error: $e(h) = [\sum_{i=1,R*k} e_i(h)]/(R*k)$
- Desviación Estándar: $S_{e(h)} = \sqrt{1/(R*(k-1)) \quad \sum_{i=1,R*k} (e_i(h) - e(h))^2}$
- Intervalo de confianza t-student: $[e(h) \pm t_{N,R*(k-1)} \quad S_{e(h)}/\sqrt{R*k}]$
- Siendo qt(0.95,36) = 1,688298 calculado con R

Siendo k = 10 y R = 4.

6 Comparativas de la estimación del error

6.1 Conjunto de datos - soybean

Algoritmo	50T	Hold Out	Hold Out Repetido	10-XV	4 x 10-XV
J48					
Error	0,411671924290221	0,103004291845494	0,109354467203066	0,0821500000000001	0,0882352941176471
Desviación	0,0195606623587688	0,012081503606538	0,0105673992752356	0,0106491783720623	0,033851824720493
Sin podar					
Error	0,416403785488959	0,107296137339056	0,113711024413243	0,09224	0,0948316283034953
Desviación	0,0195934860745905	0,0123010984913333	0,0206986099918583	0,00953312354081513	0,0350435294569074

Como podemos observar en los resultados de todos los métodos para el algoritmo J48, podemos decir que parece que el mejor método de clasificación es la validación cruzada 10 particiones pues es con el que obtenemos menor tasa de error de todos, a la par con el validación cruzada repetida, con el que obtenemos resultados parecidos.

Con el Algoritmo unprunned, mismo racionamiento, obtenemos peores resultados que con el algoritmo J48 y decimos que el mejor método de clasificación es el 10 particiones.

6.2 Conjunto de datos - vote

Algoritmo	50T	Hold Out	Hold Out Repetido	10-XV	4 x 10-XV
J48					
Error	0,077720207253886	0,0337837837837838	0,0541574738003309	0,0342699999999999	0,0351083509513742
Desviación	0,0136448245874235	0,00920790734958442	0,0146614150119994	0,0057205380477325	0,0269968624124527
Sin podar					
Error	0,077720207253886	0,0337837837837838	0,0507790954219526	0,0423699999999999	0,0403012684989429
Desviación	0,0136448245874235	0,00920790734958442	0,0118236710231369	0,00881980725412975	0,0315414221939052

Con el conjunto de datos vote, con el algoritmo J48, parece que el método de clasificación que menor tasa de error tiene sigue siendo el validación cruzada 10 particiones, seguido del repetido y el Hold Out.

Para el algoritmo unprunned, el método que mejor clasifica parece ser el Hold Out, seguido del 10-XV.

7 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

8 Preguntas sobre validación cruzada

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)

8.1 ¿Qué tasa de error se obtendría con el método 2?

Con el método 2, puede que obtengamos peores tasas de error, es decir, más altas, puesto que estamos clasificando con un conjunto de datos nuevos, no con el que hemos entrenado, por ello obtendríamos peores resultados.

8.2 ¿Cómo espera que varíe la estimación de la varianza e intervalos de confianza con el método 2 frente al método 1?

La estimación de la varianza será menos precisa puesto que estamos partiendo el conjunto de datos de manera que tenemos menos instancias para dicha estimación que destinamos para la clasificación.
En cuanto a los intervalos de confianza, mismo comportamiento, obtendríamos peores estimaciones, es decir, intervalos menos precisos, debido a la reducción de instancias en el conjunto de datos.