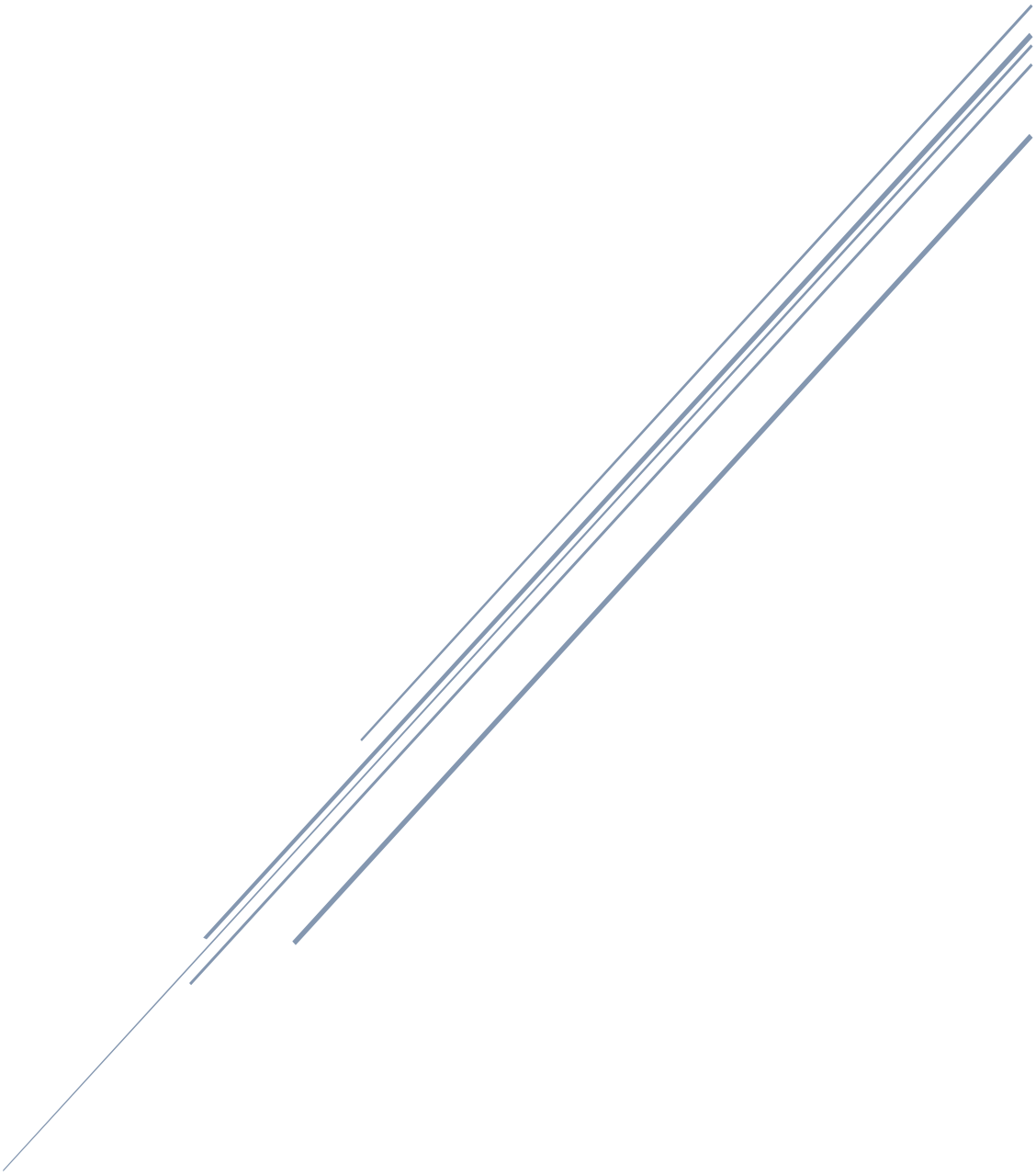
MEMORIA PRACTICA 4

TECNICAS DE APRENDIZAJE AUTOMÁTICO



CURSO 2022/23

JAVIER ABAD HERNÁNDEZ

1. **Descripción de los conjuntos de datos:**

El conjunto de datos Soybean contiene información sobre la presencia o ausencia de varias enfermedades en plantas de soja. Tiene un total de 683 instancias y 36 atributos, de los cuales 35 son atributos predictivos y uno es la clase objetivo. La clase objetivo consta de 19 posibles valores distintos, que representan diferentes enfermedades. El conjunto de datos se puede descargar de https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)

Por otro lado, el conjunto de datos Vote contiene información sobre los registros de votación de los miembros del Congreso de EE. UU. sobre ciertas cuestiones. Tiene 435 instancias y 17 atributos, de los cuales 16 son atributos predictivos y uno es la clase objetivo. La clase objetivo consta de dos posibles valores, que representan la afiliación partidista del miembro del Congreso. El conjunto de datos se puede descargar de https://archive.ics.uci.edu/ml/datasets/congressional+voting+records.

2. **Descripción de los algoritmos:**

En este trabajo se utilizarán dos algoritmos de árboles de decisión: J48 y un árbol sin podar. J48 es una implementación del algoritmo C4.5 en Weka, que construye árboles de decisión a partir de datos de entrenamiento. Un árbol sin podar, por otro lado, es un árbol que no se ha limitado en tamaño ni se ha eliminado ninguna rama durante su construcción. Estos árboles son más propensos a sobre ajustarse a los datos de entrenamiento, pero pueden proporcionar resultados más precisos en algunos casos.

Para llevar a cabo los experimentos se utilizará la herramienta Weka, una plataforma de minería de datos y aprendizaje automático que proporciona una variedad de algoritmos de clasificación y evaluación. Se utilizará Weka Explorer, una interfaz gráfica de usuario de Weka, para cargar los datos, ejecutar los algoritmos y realizar la evaluación de los resultados.

3. **Experimentos con las muestras de 50 instancias de cada conjunto de datos**

Datos	**Algoritmo**	**Método: 50 T, resto**		
		Tasa error	**Desviación estándar**	**Intervalos**
Soybean_50	J48	0,50237	0,01987	(0,4825 – 0,5222)
	Sin podar	0,50237	0,01987	(0,4825 – 0,5222)
Vote_50	J48	0,07013	0,01301	(0,0446 – 0,09569)
	Sin podar	0,07013	0,01301	(0,0446 – 0,09569)

4. **Experimentos de hold out sin repetición:**

Datos	Algoritmo	Método: Hold out		
		Tasa error	Desviación estándar	Intervalos
Soybean_50	J48	0,103	0,019913	[0,06397 - 0,14202]
	Sin podar	0, 107296	0,0202	[0,067556 - 0,14703]
Vote_50	J48	0,0337	0,01483	[0,00462 - 0,06277]
	Sin podar	0,0337	0,01483	[0,00462 - 0,06277]

Datos	Algoritmo	Tasa de error		
		2	3	4
Soybean	J48	0,09787	0,11965	0,11688
	Sin podar	0,09787	0,11965	0,12554
Vote	J48	0,06756	0,05405	0,061224
	Sin podar	0,05405	0,05405	0,061224

5. **Experimentos de hold out con repetición:**

Datos	Algoritmo	Método: Hold out repetido		
		Tasa error	Desviación estándar	Intervalos
Soybean	J48	0,10935	0,02043	[0,06929 - 0,1494]
	Sin podar	0,11259	0,02069	[0,07202 - 0,15315]
Vote	J48	0,05415	0,01863	[0,01762 - 0,090672]
	Sin podar	0,05077	0,01807	[0,01534 - 0,08661]

6. **Experimentos de validación cruzada sin repetición:**

Datos	Algoritmo	Método: 10 XV		
		Tasa error	Desviación estándar	Intervalos
Soybean	J48	0,08491	0,024731342	[0,06721 - 0,1026]
	Sin podar	0,086445	0,02734	[0,06688 - 0,10601]
Vote	J48	0,03668	0,03424	[0,06117 - 0,06117]
	Sin podar	0,03678	0,03442	[0,01216 - 0,061412]

Datos	Algoritmo	Tasa de error		
		2	3	4
Soybean	J48	0,09809	0,09077	0,07906
	Sin podar	0,10541347	0,102557545	0,089322251
Vote	J48	0,0322	0,0369	0,034513
	Sin podar	0,04365	0,04159	0,03916

7. **Experimentos de validación cruzada con repetición**

Datos	Algoritmo	Método: Validación cruzada repetida		
		Tasa error	Desviación estándar	Intervalos
Soybean	J48	0,08823	0,033851825	[0,06721 - 0,09906]
	Sin podar	0,09593	0,03428	[0,077408 - 0,10689]
Vote	J48	0,0351	0,02699	[0,02647 - 0,04374]
	Sin podar	0,0403	0,03154	[0,0302– 0,05038]

8. Tablas comparativas y discusión de resultados

TABLA RESUMEN: soybean

Algoritmo	50 instan. entrenam.	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48					
Error	0,50237	0,103	0,10935	0,08491	0,08823
Desviación	0,01987	0,019913	0,02043	0,024731342	0,033851825
Intervalos	(0,4825 – 0,5222)	[0,06397 - 0,14202]	[0,06929 - 0,1494]	[0,06721 - 0,1026]	[0,06721 - 0,09906]
Sin podar					
Error	0,50237	0, 107296	0,11259	0,086445	0,09593
Desviación	0,01987	0,0202	0,02069	0,02734	0,03428
Intervalos	(0,4825 – 0,5222)	[0,067556 - 0,14703]	[0,07202 - 0,15315]	[0,06688 - 0,10601]	[0,077408 - 0,10689]

TABLA RESUMEN: vote

Algoritmo	50 instan. entrenam.	Hold out	Hold out repetido (4)	10-XV	4 x 10-XV
J48					
Error	0,07013	0,0337	0,05415	0,03668	0,0351
Desviación	0,0130	0,01483	0,01863	0,03424	0,02699
Intervalos	(0,0446 – 0,09569)	[0,00462 - 0,06277]	[0,01762 - 0,090672]	[0,06117 - 0,06117]	[0,02647 - 0,04374]
Sin podar					
Error	0,07013	0,0337	0,05077	0,03678	0,0403
Desviación	0,0130	0,01483	0,01807	0,03442	0,03154
Intervalos	(0,0446 – 0,09569)	[0,00462 - 0,06277]	[0,01534 - 0,08661]	[0,01216 - 0,061412]	[0,0302– 0,05038]

Los algoritmos han ido de mayor a menor error, siendo el 50T instancias el que tiene mayor tasa de error, seguido por el Hold Out y la validación cruzada.

Esta situación tiene bastante sentido, ya que, al entrenar con menos instancias, crea un peor modelo y por tanto mayor tasa de error.

Por otro lado, podemos ver que, a menor tasa de error, mayor desviación. Esto compensa para que los intervalos se mantengan “estables”.

Por último, hay que mencionar que la validación cruzada es la que presenta menor tasa de errores ya que es la que se entrena con un mayor conjunto de datos, y da menor lugar a fallo.

9. Preguntas sobre validación cruzada

- ¿Qué tasa de error se obtendría con el método 2?
  Se debería obtener la misma tasa de error.
- ¿Cómo espera que varíe la estimación de la varianza con el método 2 frente al método 1?
  Se espera que el valor sea inferior.
- ¿Y los intervalos de confianza?
  Al esperar una menor varianza, la desviación será menor y por ende, se esperan unos intervalos más pequeños.

10. Referencias:

- Apuntes de Técnicas de Aprendizaje Automático, 2022/23.