Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)
Medidas de evaluación en Árboles de decisión

1

Profesora: *Dra. Fabiola Ocampo Botello*

2

Un algoritmo de inducción, o más concisamente un **inductor** (también conocido como aprendiz), es una entidad que obtiene un conjunto de entrenamiento y forma un modelo que generaliza la relación entre los atributos de entrada y el atributo objetivo. Por ejemplo, un inductor puede tomar como entrada tuplas de entrenamiento específicas con la etiqueta de clase correspondiente y producir un clasificador (Rokach, L. & Maimon, O., 2015).

Los inductores de árboles de decisión son algoritmos que construyen automáticamente un árbol de decisiones a partir de un conjunto de datos determinado (Rokach, L. & Maimon, O., 2015).

Rokach, L. & Maimon, O. (2015) establecen que la meta de un algoritmo de clasificación se puede definir formalmente como:

Dado un conjunto de entrenamiento S con atributos de entrada
A = {a1, a2, . . . , an}
y un atributo nominal y y una distribución desconocida D, la meta es inducir un clasificador óptimo con el mínimo error de generalización.

Notación:

DT	Representa el inductor del árbol de decisión.
DT(S)	Representa un árbol de clasificación que se generó al ejecutar DT sobre el conjunto de datos S.
DT(S)(x_q)	Es la predicción de x_q usando DT(S).

3

Rokach, L. & Maimon, O. (2015) establecen que el **error de generalización** es definido como la tasa de clasificación errónea sobre la distribución D, en caso de atributos nominales puede ser expresado como:

$$\varepsilon(DT(S), D) = \sum_{(x,y)\in U} D(x,y) \cdot L(y, DT(S)(x)), \qquad (3.1)$$

Donde $L(y, DT(S)(x))$ es una función cero o uno definido como:

$$L(y, DT(S)(x)) = \begin{cases} 0 & if\ y = DT(S)(x) \\ 1 & if\ y \neq DT(S)(x) \end{cases}. \qquad (3.2)$$

Imágenes tomadas de Rokach, L. & Maimon, O. (2015)

4

Evaluación de árboles de clasificación

Rokach, L. & Maimon, O. (2015) establecen los siguientes aspectos de la evaluación de los árboles de clasificación.

La **exactitud de la clasificación** se expresa como uno menos el error de generalización.

El **error de entrenamiento** es definido como el porcentaje de ejemplos en el conjunto de entrenamiento que fueron correctamente clasificados en el árbol de clasificación, lo cual se expresa:

$$\hat{\varepsilon}(DT(S), S) = \sum_{(x,y)\in S} L(y, DT(S)(x)), \qquad (4.1)$$

5

Donde $L(y, DT(S)(x))$ se define igual que la expresión 3.2.

Existen dos formas de estimar el error de generalización:
1. Teórico
2. Empírico

6

Estimación empírica del error de generalización

Uno de los enfoques para estimar el error de generalización es el método de retención (*holdout method*) en el que el conjunto de datos dado se divide aleatoriamente en dos conjuntos: Conjuntos de entrenamiento y prueba (Rokach, L. & Maimon, O, 2015) .

Por lo general, dos tercios de los datos se consideran para el conjunto de entrenamiento y los datos restantes se asignan al conjunto de prueba.

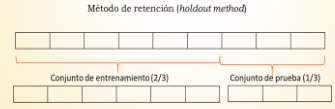

Imagen: Elaboración propia

El submuestreo aleatorio (*Random subsampling*) y la validación cruzada n-fold (*n-fold cross-validation*) son dos métodos comunes de remuestreo (Rokach, L. & Maimon, O, 2015):

- En el submuestreo aleatorio, los datos se dividen aleatoriamente varias veces en conjuntos de entrenamiento y pruebas disjuntos. Los errores obtenidos de cada partición se promedian.
- En la validación cruzada n-fold, los datos se dividen aleatoriamente en n subconjuntos mutuamente excluyentes de aproximadamente el mismo tamaño. Un inductor es entrenado y probado n veces; cada vez se prueba en uno de los k pliegues (fold) y se entrena utilizando los n-1 pliegues (fold) restantes.

7

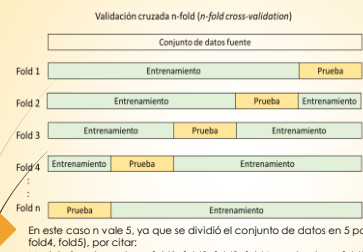

Imagen: Elaboración propia

8

En este caso n vale 5, ya que se dividió el conjunto de datos en 5 partes (fold1, fold2, fold3, fold4, fold5), por citar:
Modelo 1: entrenado en fold1+fold2+fold3+fold4 y probado en fold5
Modelo 2: entrenado en fold1+fold2+fold3+fold5 y probado en fold4
Modelo 3: entrenado en fold1+fold2+fold4+fold5 y probado en fold3
Modelo 4: entrenado en fold1+fold3+fold4+fold5 y probado en fold2
Modelo 5: entrenado en fold2+fold3+fold4+fold5 y probado en fold1

9

Rokach, L. & Maimon, O. (2015:31-60) presentan los siguientes medidas de evaluación en los árboles de decisión.

La **precisión** (*Accuracy*) no es una medida suficiente para evaluar un modelo con una distribución desbalanceada de las clases.

La **sensibilidad** (*Sensitivity*) (también conocida como recuerdo (*recall*)) evalúa qué tan bien el clasificador puede reconocer muestras positivas y se define como:

$$Sensitivity = \frac{true\_positive}{positive}$$

Donde *true_positive* corresponde al número de muestras positivas verdaderas y *positive* es el número de muestras positivas.

La medida de especificidad (*specificity*) mide que tan bien el clasificador puede reconocer las muestras negativas. Se define como (Rokach, L. & Maimon, O.,2015) :

$$Specificity = \frac{true\_negative}{negative}$$

Donde *true_negative* corresponde al número de ejemplos de verdaderos negativos y *negative* al número de muestras negativas.

10

Otra medida se llama precisión (*Precision*). La precisión (*Precision*) mide cuántos ejemplos clasificados como clase "positiva" son realmente "positivos". Esta medida es útil para evaluar clasificadores nítidos que se utilizan para clasificar un conjunto de datos completo. Formalmente (Rokach, L. & Maimon, O.,2015):

$$Precision = \frac{true\_positive}{true_{positive} + false\_positive}$$

Matriz de confusión

La evaluación del desempeño de un modelo de clasificación considera dos aspectos:

1. La cantidad de registros previstos por el modelo de forma adecuada.
2. La cantidad de registros previstos por el modelo de forma inadecuada.

Lo anterior se presenta en una **matriz de confusión**.

Ejemplo de una matriz de confusión:

		Clase prevista	
		Clase 1	Clase 0
Clase actual	Clase 1	f11	f10
	Clase 0	f01	f00

11

Precisión (*Accuracy*) = $\frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones}$ Precisión = $\frac{f11+f00}{f11+f10+f01+f00}$

Tasa de erros (*Error rate*) = $\frac{Número\ de\ predicciones\ incorrectas}{Número\ total\ de\ predicciones}$ Error rate = $\frac{f10+f01}{f11+f10+f01+f00}$

Table 4.1 A confusion matrix.

	Predicted negative	Predicted positive
Negative Examples	A	B
Positive Examples	C	D

Rokach, L. & Maimon, O. (2015:37).

- Accuracy is: $(a+d)/(a+b+c+d)$
- Misclassification rate is: $(b+c)/(a+b+c+d)$
- Precision is: $d/(b+d)$
- True positive rate (Recall) is: $d/(c+d)$
- False positive rate is: $b/(a+b)$
- True negative rate (Specificity) is: $a/(a+b)$
- False negative rate is: $c/(c+d)$

12

Imágenes tomadas de Rokach, L. & Maimon, O. (2015:37).

Ejemplos de la Matriz de Confusión

Ejercicio adaptado del libro de: Bennet, Briggs & Triola (2011). Razonamiento estadístico. Pearson. México.

Suponga que usted es un doctor o una doctora y tiene una paciente con un tumor en el pecho. Usted sabe que 1 de cada 100 tumores es maligno.

Un estudio de mamografía revela un *resultado positivo*.

Usted sabe que la mamografía (que tiene un 85% de precisión) reporta positivo cuando el tumor es maligno.

¿Esto significa que tiene cáncer de pecho?

¿Le dirá a la paciente que se prepare para un tratamiento contra el cáncer?

Suponga lo siguiente:

Se realizaron 10,000 mamografías a mujeres con tumores. Suponga que el 1% de los tumores son malignos

10,000 * 0.01 = 100 tumores con cáncer.

9,900 de los tumores no son cancerígenos.

13

Analice lo siguiente:

	El tumor es maligno	El tumor es benigno	Total
Mamografía positiva	85 Positivo verdadero	1,485 Positivo falso	1,570
Mamografía negativa	15 Negativo falso	8,415 Negativo verdadero	8,430
Total	100	9,900	10,000

Resumen de resultados para 10,000 mamografías (cuando en realidad 100 tumores son malignos y 9900 son benignos)

El número total de positivos fue:

85 (positivos verdaderos) que realmente tienen cáncer, y

1,485 (positivos falsos) que realmente no tiene cáncer.

Dando un total de: 1,570.

La probabilidad de que un resultado positivo en realidad signifique cáncer es de 85/1570 = 0.054% o 5.4%.

14

Comprensibilidad

Rokach, L. & Maimon, O. (2015) establece que El criterio de comprensibilidad (también conocido como interpretabilidad) se refiere a qué tan bien los humanos captan el clasificador inducido. Mientras que el error de generalización mide cómo el clasificador se ajusta a los datos, la comprensibilidad mide el "ajuste mental" de ese clasificador.

Para dominios como el diagnóstico médico, los usuarios deben comprender cómo el sistema toma sus decisiones para estar seguros del resultado.

Rokach, L. & Maimon, O. (2015) expresan que La comprensibilidad puede variar entre diferentes clasificadores creados por el mismo inductor. Por ejemplo, en el caso de los árboles de decisión, el tamaño (número de nodos) de los árboles inducidos también es importante. Se prefieren los árboles más pequeños porque son más fáciles de interpretar.

15

De acuerdo con un principio fundamental en la ciencia, conocido como la **navaja de Occam**, cuando se busca la explicación de cualquier fenómeno, uno debe hacer la menor cantidad posible de suposiciones y eliminar aquellas que no hacen ninguna diferencia en las predicciones observables de la hipótesis explicativa.

16

Robustez

Rokach, L. & Maimon, O. (2015) expresan que la capacidad del modelo para manejar el ruido o los datos con valores perdidos y hacer predicciones correctas se llama robustez. Además,

- Los diferentes algoritmos de árboles de decisión tienen diferentes niveles de robustez.
- Para estimar la robustez de un árbol de clasificación, es común entrenar el árbol en un conjunto de entrenamiento limpio y luego entrenar un árbol diferente en un conjunto de entrenamiento ruidoso.
- El conjunto de entrenamiento ruidoso suele ser el conjunto de entrenamiento limpio al que se han agregado algunas instancias ruidosas artificiales. El nivel de robustez se mide como la diferencia en la precisión de estas dos situaciones.

17

Estabilidad

Rokach, L. & Maimon, O. (2015) expresan que

Formalmente, la estabilidad de un algoritmo de clasificación se define como el grado en que un algoritmo genera resultados repetibles, dados diferentes lotes de datos del mismo proceso. Los usuarios ven el algoritmo de aprendizaje como un oráculo. Obviamente, es difícil confiar en un oráculo que dice algo radicalmente diferente cada vez que realiza un ligero cambio en los datos.

18

Sobreajuste y Subajuste

Rokach, L. & Maimon, O. (2015) mencionan que El concepto de sobreajuste es muy importante en la minería de datos. Se refiere a la situación en la que el algoritmo de inducción genera un clasificador que se ajusta perfectamente a los datos de entrenamiento pero ha perdido la capacidad de generalizar a instancias no presentadas durante el entrenamiento. En otras palabras, en lugar de aprender, el clasificador simplemente memoriza las instancias de entrenamiento. El sobreajuste se reconoce generalmente como una violación del principio de la navaja de Occam.

19

Imagen tomada de la página web del Ing. Juan Ignacio Bagnato.

Dirección Web:

https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/

20

Escalabilidad a grandes bases de datos (Scalability to Large Datasets)

Rokach, L. & Maimon, O. (2015) mencionan que La escalabilidad se refiere a la capacidad del método para construir el modelo de clasificación de manera eficiente dada una gran cantidad de datos.

Los enfoques para tratar con una gran cantidad de registros incluyen:

- Métodos de muestreo: los estadísticos seleccionan registros de una población mediante diferentes técnicas de muestreo.
- Agregación: reduce el número de registros al tratar un grupo de registros como uno o al ignorar los subconjuntos de registros "sin importancia".
- Procesamiento masivo en paralelo.
- Métodos de almacenamiento eficientes: permiten que el algoritmo maneje muchos registros. Por ejemplo una estructura de datos de lista de atributos.
- Reducción del espacio de búsqueda del algoritmo: por ejemplo, el algoritmo PUBLIC [Rastogi y Shim (2000)] integra el crecimiento y la poda de los árboles de decisión mediante el uso del enfoque de Longitud mínima de descripción (*Minimum Description Length, MDL*) para reducir la complejidad computacional.

Referencias bibliográficas

Bagnato, Juan Ignacio. (2017). Aprende Machine Learning. Portal Web. Disponible en: https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/

Bennet, Briggs & Triola (2011). *Razonamiento estadístico*. Pearson. México.

Rokach, L. & Maimon, O. (2015). Data Mining with decision trees. Theory and Applications. Second Edition. World Scientific Publishing Co. Pte. Ltd.