PrÁctica 7

NAIVE BAYES

TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

JAVIER ABAD HERNÁNDEZ

**1. Dadas dos variables aleatorias discretas, X e Y, y dada su distribución de probabilidad conjunta que aparece en la tabla, se pide:**

**a. ¿Cumple la distribución conjunta las propiedades de una distribución de probabilidades?**

Para verificar si la distribución conjunta cumple con las propiedades de una distribución de probabilidad, es necesario comprobar que la suma de todas las probabilidades sea igual a 1 y que todas las probabilidades sean no negativas.

La suma de todas las probabilidades es como vemos en la tabla, = 1:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | x1 | x2 | x3 | x4 | P(Y) |
| y1 | 2/16 | 1/16 | 1/16 | 1/16 | 5/16 |
| y2 | 1/16 | 2/16 | 2/16 | 1/16 | 6/16 |
| y3 | 1/16 | 1/16 | 1/16 | 0 | 3/16 |
| y4 | 0 | 2/16 | 0 | 0 | 2/16 |
| P(X) | 1/4 | 6/16 | 1/4 | 2/16 | 1,00 |

Todas las probabilidades son no negativas, por lo tanto, la distribución conjunta cumple con las propiedades de una distribución de probabilidad.

**b. ¿Cuál es la probabilidad de 𝑃(𝑋=x1)?**

La probabilidad de P(X=x1) se puede calcular sumando todas las probabilidades en la fila correspondiente a x1:

P(X=x1) = P(X=x1, Y=y1) + P(X=x1, Y=y2) + P(X=x1, Y=y3) + P(X=x1, Y=y4) =

= 2/16 + 1/16 + 1/16 + 0 = 1/4

Por lo tanto, la probabilidad de P(X=x1) es 1/4.

**c. ¿Cuáles son las distribuciones marginales de cada 𝑃(𝑋=x) y P(Y=y)?**

**i. Distribución marginal: distribución de probabilidad sobre un subconjunto de las variables aleatorias del espacio probabilístico**

Para calcular las distribuciones marginales de X e Y, es necesario sumar las probabilidades en cada fila y en cada columna, respectivamente. Esto ya lo hemos realizado en la tabla mostrada en el apartado a)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | x1 | x2 | x3 | x4 | P(Y) |
| y1 | 2/16 | 1/16 | 1/16 | 1/16 | 5/16 |
| y2 | 1/16 | 2/16 | 2/16 | 1/16 | 6/16 |
| y3 | 1/16 | 1/16 | 1/16 | 0 | 3/16 |
| y4 | 0 | 2/16 | 0 | 0 | 2/16 |
| P(X) | 1/4 | 6/16 | 1/4 | 2/16 | 1,00 |

1. La distribución marginal es la distribución de probabilidad de una variable aleatoria en el espacio probabilístico, sin tener en cuenta las demás variables. En este caso, las distribuciones marginales son P(X) y P(Y), respectivamente.

**d. ¿Verifican las distribuciones marginales las propiedades de una distribución de probabilidades?**

Las distribuciones marginales también deben cumplir con las propiedades de una distribución de probabilidad, es decir, la suma de todas las probabilidades debe ser igual a 1 y todas las probabilidades deben ser no negativas. En este caso, se puede verificar que ambas distribuciones marginales cumplen con estas propiedades.

**2. Utilizando el conjunto de datos weather.nominal.practica que se proporciona, determinar la clasificación Naive Bayes de las siguientes instancias, utilizando la estimación de máxima verosimilitud (frecuencial) y sin utilizar ninguna herramienta de minería de datos:**

**a. 𝑋1 < sunny, cool, normal, false >**

Contamos las instancias que hay para cada posibilidad y hallamos su probabilidad:

Clases:

*P(Play = no) = 5/14 = 0,357*

*P(Play = yes) = 9/14 = 0,643*

Atributos:

*P(Outlook = sunny | Play = no) = 3/5 = 0,6*

*P(Outlook = sunny | Play = yes) = 2/9 = 0,22*

*P(Temperature = cool | Play = no) = 1/5 = 0,2*

*P(Temperature = cool | Play = yes) = 5/9 = 0,55*

*P(Humidity = normal | Play = no) = 1/5 = 0,2*

*P(Humidity = normal | Play = yes) = 6/9 = 0,66*

*P(Windy = false | Play = no) = 2/5 = 0,4*

*P(Windy = false | Play = yes) = 5/9 = 0,55*

*P(Play = no | <sunny, cool, normal, false>)= P(Play = no) \* P(Outlook = sunny | Play = no) \* P(Temperature = cool | Play = no) \* P(Humidity = normal | Play = no) \* P(Windy = false | Play = no) = 0,357 \* 0,6 \* 0,2 \*0,2 \* 0,4 = 0,0034272*

*P(Play = yes | <sunny, cool, normal, false>)= P(Play = yes) \* P(Outlook = sunny | Play = yes) \* P(Temperature = cool | Play = yes) \* P(Humidity = normal | Play = yes) \* P(Windy = false | Play = yes) = 0,643 \* 0,22 \* 0,55 \* 0,66 \* 0,55 = 0,0282424*

Vnb = yes, ya que tiene una probabilidad mayor

**b. 𝑋2 =< overcast, hot, high, true >**

Clases

*P(Play = no) = 5/14 = 0,357*

*P(Play = yes) = 9/14 = 0,643*

Atributos

*P(Outlook = overcast | Play = no) = 0/5 = 0*

*P(Outlook = overcast | Play = yes) = 2/9 = 0,22*

*P(Temperature = hot | Play = no) = 2/5 = 0,4*

*P(Temperature = hot | Play = yes) = 0/9 = 0*

*P(Humidity = high | Play = no) = 4/5 = 0,8*

*P(Humidity = high | Play = yes) = 3/9 = 0,33*

*P(Windy = true | Play = no) = 3/5 = 0,6*

*P(Windy = true | Play = yes) = 4/9 = 0,44*

*P(Play = no | <overcast, hot, high, true>)= P(Play = no) \* P(Outlook = overcast | Play = no) \* P(Temperature = hot | Play = no) \* P(Humidity = high | Play = no) \* P(Windy = true | Play = no) = 0,357 \* 0 \* 0,4 \* 0,8 \* 0,6 = 0*

*P(Play = yes | <overcast, hot, high, true>)= P(Play = yes) \* P(Outlook = overcast | Play = yes) \* P(Temperature = hot | Play = yes) \* P(Humidity = high | Play = yes) \* P(Windy = true | Play = yes) = 0,643\* 0,22 \* 0 \* 0,33 \* 0,44 = 0*

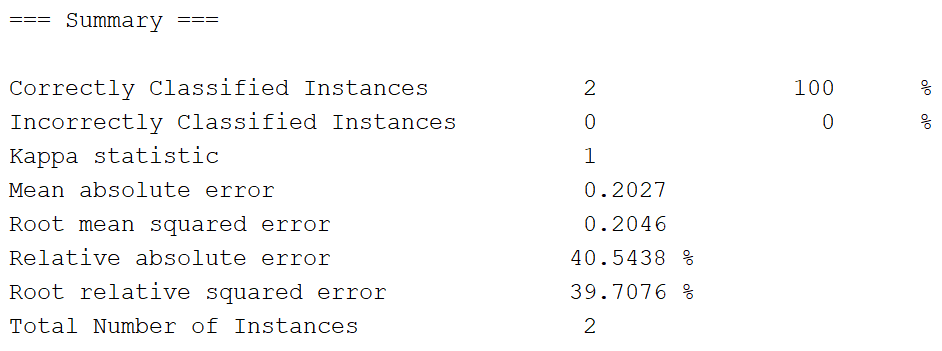
Como tenemos valores que son = 0 (outlook y temperature) , voy a proceder a realizar una corrección de muestreo, donde ignoramos ambos.

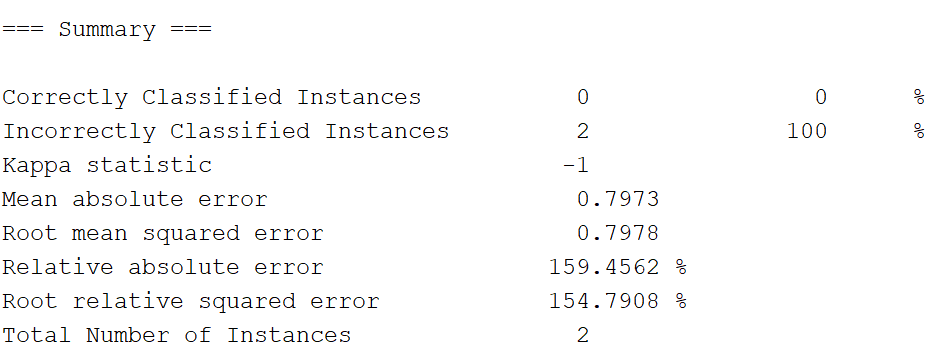
no = 0,357 \* 0,4 \* 0,8 \* 0,6 = 0,068544

yes = 0,643 \* 0,22 \* 0,33 \* 0,44 = 0,020539

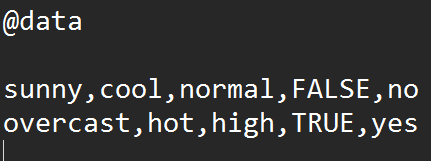
Vnb = no, ya que tiene una probabilidad mayor.

**3. Utilizando Weka y el clasificador NaiveBayes determinar la clasificación de los ejemplos anteriores: a. ¿Coindice con la clasificación calculada en el ejercicio anterior?**





Esto ocurre por este conjunto de datos de prueba:



**4. Entrenar, con Weka, un clasificador Naive Bayes para el conjunto de datos weather.nominal.practica**

**a. Estimar la tasa de error cometida por el clasificador utilizando validación cruzada de 10 particiones.**

**Interfaz de usuario gráfica, Texto

Descripción generada automáticamente**

Como podemos ver, la tasa de error con validación cruzada de 10 particiones para Naive Bayes, es de 0.428571, es decir, el 42,8571%.

**b. Examinar la salida proporcionada por el Explorer y determinar cómo está estimando esta implementación de Naive Bayes los parámetros del clasificador.**

**Imagen que contiene Tabla

Descripción generada automáticamente**Weka utiliza estimación Bayesiana, se resuelve con esta ecuación:

P(A = ai | B = bj) = (nc+mp)/(n+m)

Usando el estimador de Laplace:

p = 1 / Val(ai) y m = val(ai)

**5. El conjunto de datos weather.nominalX6 se ha generado repitiendo cada instancia del conjunto weather.nominal.practica seis veces. Entrenar con Weka un clasificador Naive Bayes para este conjunto de datos:**

**a. Estimar la tasa de error cometida por el clasificador utilizando validación cruzada de 10 particiones.**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Como podemos ver, la tasa de error es del 15.4762% es decir, es 0.154762.

**b. Compare esta tasa de error con la estimada en el ejercicio anterior y discuta los resultados.**

La tasa de error en 5a es significativamente menor que en 4a. Esto podría deberse a que al repetir instancias en el conjunto de datos weather.nominalX6, se introdujo un sesgo en los datos que permitió al clasificador aprender a reconocer instancias específicas en lugar de generalizar a nuevos datos. Esto puede resultar en una mayor precisión en los datos de entrenamiento, pero una menor precisión en nuevos datos no vistos.