Naive Bayes Classification

LJ Brown

April 19, 2018

Homework 5, Question 5

Derivation of Bayes' Theorem from Kolmogorov's definition of Conditional Probability, and using it to guess the value of the missing entry (?) in Table 1.

Table 1

Weather	Temp	Humidity	Windy	Play
Rainy	Cool	Normal	FALSE	yes
Rainy	Cool	Normal	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Sunny	Hot	Normal	FALSE	Yes
Overcast	Mild	High	TRUE	Yes
Sunny	Mild	High	TRUE	No
Sunny	Cool	High	False	?

Conditional Probability Definition [1]

$$P\left(A|B\right) = \frac{P\left(A \cap B\right)}{P\left(B\right)} \tag{1}$$

Bayes' Theorem

$$P\left(A|B\right) = \frac{P\left(A\right)P\left(B|A\right)}{P\left(B\right)} \tag{2}$$

[1] Conditional Probability definition from Kolmogorov's probability theory.

Limitations

The Naive Bayes Classifier outlined bellow can only be applied to datasets where the column values are independent categorical variables.

Derivation of Bayes' Theorem

Bayes' Theorem can be derived from the definition of Conditional Probability in equation 1 by first substituting $A = B$ and $B = A$,

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \qquad \text{(eq 1: A = B, B = A substitution)}$$

This equation can be solved for $P(A \cap B)$:

$$P(B \cap A) = P(A)\,P(B|A)$$

$$P(A \cap B) = P(B \cap A)$$

$$P(A \cap B) = P(A)\,P(B|A) \tag{*}$$

This definition of $P(A \cap B)$ in equation * can be substituted into the equation defining Conditional Probability (equation 1) to find Bayes' Theorem (equation 2):

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)} \tag{2}$$

Question 5

Final Row

Weather	Temp	Humidity	Windy	Play
Sunny	Cool	High	False	?

Table To Equation Conversions

v_{column} = selected categorical value of specified column

$$P\left(v_{column}\right) = \frac{\text{number of rows where column entry is } v_{column}}{\text{total number of rows}}$$

$$P\left(v_{column_1} | v_{column_2}\right) = \frac{\text{number of rows where } column_1 \text{ is } v_{column_1} \text{ and } column_2 \text{ is } v_{column_2}}{\text{number of rows where } column_2 \text{ is } v_{column_2}}$$

Plugging in the given values from final row into equation variables:

$v_{Weather}$ = Sunny,
v_{Temp} = Cool,
$v_{Humidity}$ = High,
v_{Windy} = False

And (in place of the ?) to find the probability that the value in the "Play" column is "Yes"...

v_{Play} = Yes.

Values directly substituted into Bayes' Theorem:

$$P\left(A|B\right) = \frac{P\left(A\right) P\left(B|A\right)}{P\left(B\right)} \tag{2}$$

$$P\left(v_{Play}|v_{Weather} \ldots \cap v_{Windy}\right) = \frac{P\left(v_{Play}\right) P\left(v_{Weather} \ldots \cap v_{Windy}|v_{Play}\right)}{P\left(v_{Weather} \ldots \cap v_{Windy}\right)}$$

Where,

$$v_{Weather} \cap v_{Temp} \cap v_{Humidity} \cap v_{Windy} = v_{Weather} \ldots \cap v_{Windy}$$

If the columns are independent variables then,

$$P\left(v_{Weather} \ldots \cap v_{Windy}\right) = P\left(v_{Weather}\right) \ldots P\left(v_{Windy}\right) \tag{3}$$

$$P\left(v_{Weather} \ldots \cap v_{Windy} | v_{Play}\right) = P\left(v_{Weather} | v_{Play}\right) \ldots P\left(v_{Windy} | v_{Play}\right) \tag{4}$$

Final Equation

Substituting equations 3 and 4 into the version of Bayes' Theorem above:

$$P\left(v_{Play} | v_{Weather} \ldots \cap v_{Windy}\right) = \frac{P\left(v_{Play}\right) P\left(v_{Weather} | v_{Play}\right) \ldots P\left(v_{Windy} | v_{Play}\right)}{P\left(v_{Weather}\right) \ldots P\left(v_{Windy}\right)}$$

Results

$$P\left(v_{Play} | v_{Weather} \ldots \cap v_{Windy}\right) = \frac{0.6 * 0.33 * 0.5 * 0.33 * 0.83}{0.4 * 0.5 * 0.4 * 0.7} = 0.49$$