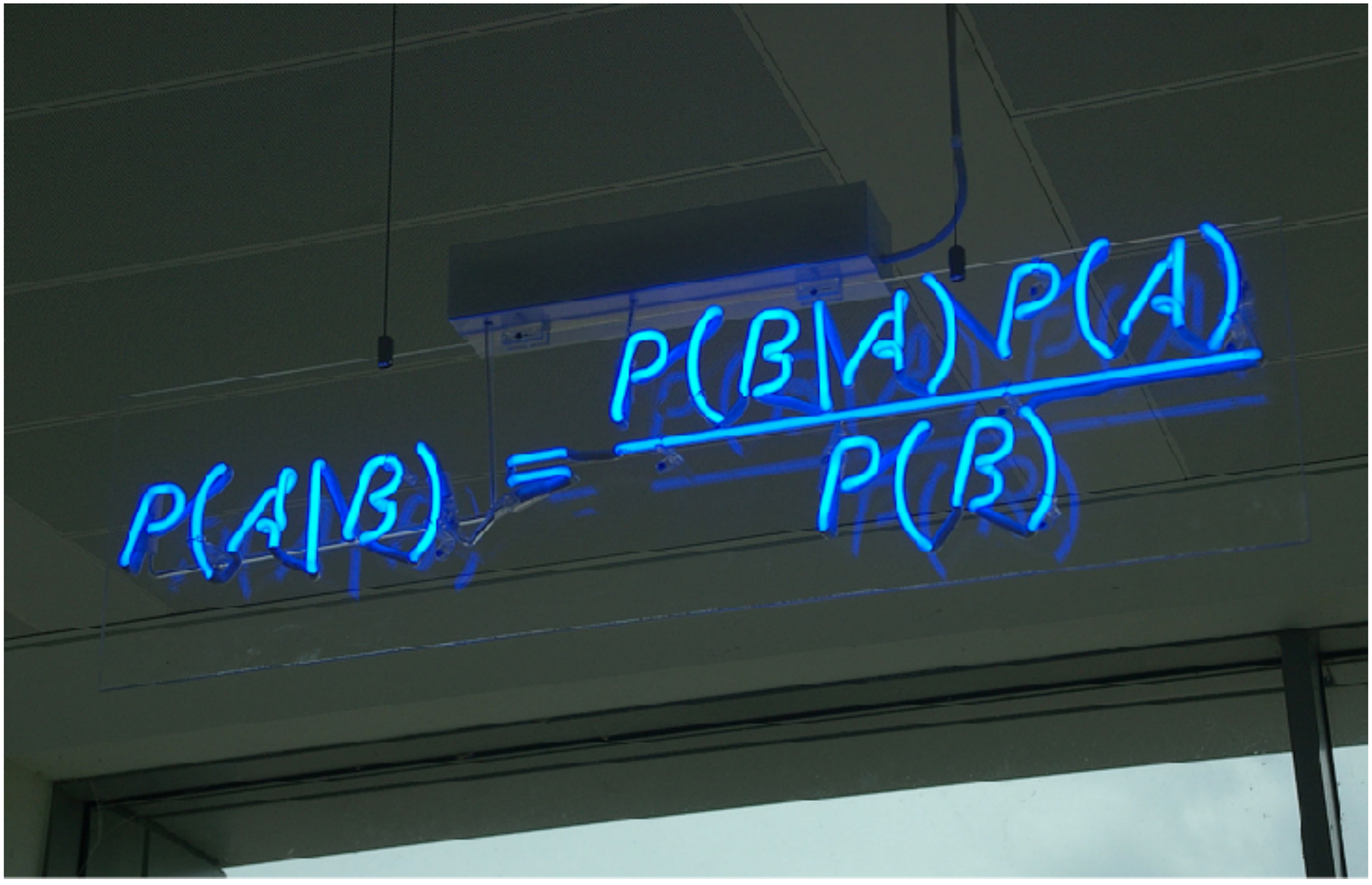

Classification and The Bayes Theorem

Leandro L. Minku

Machine Learning and Probabilities

We can assume that this process generates data based on an unknown joint probability distribution.

Data are generated from some underlying process.

Distributions on Multiple Random Variables

Conditional probability distribution of dependent variable Y given independent variables **X**

Probability distribution of dependent variable Y

$$P(\mathbf{X},Y) = P(\mathbf{X})\,P(Y|\mathbf{X}) = P(Y)\,P(\mathbf{X}|Y)$$

Joint probability distribution of independent variables **X** and dependent variable Y

Probability distribution of independent variables **X**

Conditional probability distribution of independent variables **X** given dependent variable Y

Observed Values

Probability of output class being c given that the input values are **a**

Probability of observing input values **a** with output class c

Probability of observing input values **a**

Probability of input values being **a** given that the output class is c

$$P(\mathbf{X}=\mathbf{a}, Y=c) = P(\mathbf{X}=\mathbf{a})P(Y=c|\mathbf{X}=\mathbf{a}) = P(Y=c)P(\mathbf{X}=\mathbf{a}|Y=c)$$

Probability of observing output class c

Bayes Theorem

Probability of output class being c given that the input values are **a**

Probability of observing output class c

$$P(\mathbf{a},c) = P(\mathbf{a}) \quad P(c|\mathbf{a}) = \frac{P(c) \quad P(\mathbf{a}|c)}{}$$

Probability of observing input value **a**

Probability of input values being **a** given that the output class is c

Our machine learning task can then be seen as learning probabilities, which can then be used for making predictions based on the Bayes Theorem.

Learning Probabilities

Intuitively, if we ignore the variable Wash, the probability of Cavity = yes when Pain = yes based on the data above should be larger than the probability of Cavity = yes when Pain = no.

Person	x_1 (Wash)	x_2 (Pain)	y (Cavity)
P1	no	yes	yes
P2	no	yes	yes
P3	yes	yes	yes
P4	yes	no	no
P5	yes	no	no
P6	no	no	no

We can learn probabilities by keeping track of the frequencies associated to different input and output values.

Illustrative Example for One Independent Variable

Training Set

Person	x_1 (Wash)	y (Cavity)
P1	no	yes
P2	no	yes
P3	yes	yes
P4	yes	no
P5	yes	no
P6	no	no

Model

Frequency Table	Cavity = no	Cavity = yes	Total:
Wash = no	1	2	3
Wash = yes	2	1	3
Total:	3	3	6

Illustrative Example for One Independent Variable

Training Set

Person	x_1 (Wash)	y (Cavity)
P1	no	yes
P2	no	yes
P3	yes	yes
P4	yes	no
P5	yes	no
P6	no	no

P(Wash=no|Cavity=no) = 1 / 3

P(Wash=yes|Cavity=no) = 2 / 3

P(Wash=no|Cavity=yes) = 2 / 3

P(Wash=yes|Cavity=yes) = 1 / 3

Model

Frequency Table	Cavity = no	Cavity = yes	Total:
Wash = no	1	2	3
Wash = yes	2	1	3
Total:	3	3	6

P(Cavity=no) = 3 / 6

P(Cavity=yes) = 3 / 6

P(Wash=no) = 3 / 6

P(Wash=yes) = 3 / 6

How to Make Predictions For An Example With Input Value a?

$$P(c|a) = \frac{P(c)\,P(a|c)}{P(a)}$$

Apply Bayes Theorem: Calculate P(c|a) for each class c and then predict the class associated to the maximum P(c|a).

Example of Applying Bayes Theorem

Model

Frequency Table	Cavity = no	Cavity = yes	Total:
Wash = no	1	2	3
Wash = yes	2	1	3
Total:	3	3	6

New example: (Wash=no, y=?)

$$P(c|a) = \frac{P(c)\,P(a|c)}{P(a)}$$

Example of Applying Bayes Theorem

Model

Frequency Table	Cavity = no	Cavity = yes	Total:
Wash = no	1	2	3
Wash = yes	2	1	3
Total:	3	3	6

New example: (Wash=no, y=?)

$$P(c|a) = \frac{P(c)\,P(a|c)}{P(a)}$$

P(Cavity=no | Wash=no) = ?

P(Cavity=yes | Wash=no) = ?

Example of Applying Bayes Theorem

Model

Frequency Table	Cavity = no	Cavity = yes	Total:
Wash = no	1	2	3
Wash = yes	2	1	3
Total:	3	3	6

New example: (Wash=no, y=?)

$$P(c|a) = \frac{P(c)\,P(a|c)}{P(a)}$$

$$P(\text{Cavity=no} \mid \text{Wash=no}) = \frac{P(\text{Cavity=no})\,P(\text{Wash=no} \mid \text{Cavity=no})}{P(\text{Wash=no})} = \frac{3/6 * \ 1/3}{3/6} = 0.33$$

Example of Applying Bayes Theorem

Model

Frequency Table	Cavity = no	Cavity = yes	Total:
Wash = no	1	2	3
Wash = yes	2	1	3
Total:	3	3	6

New example: (Wash=no, y=?)

$$P(c|a) = \frac{P(c)\,P(a|c)}{P(a)}$$

$$P(\text{Cavity=yes} \mid \text{Wash=no}) = \frac{P(\text{Cavity=yes})P(\text{Wash=no} \mid \text{Cavity=yes})}{P(\text{Wash=no})} = \frac{3/6 * 2/3}{3/6} = 0.67$$

Example of Prediction Based on the Bayes Theorem

New example: (Wash=no, y=?)

$$P(\text{Cavity=no} \mid \text{Wash=no}) = \frac{P(\text{Cavity=no})\,P(\text{Wash=no} \mid \text{Cavity=no})}{P(\text{Wash=no})} = \frac{3/6 * 1/3}{3/6} = 0.33$$

$$P(\text{Cavity=yes} \mid \text{Wash=no}) = \frac{P(\text{Cavity=yes})P(\text{Wash=no} \mid \text{Cavity=yes})}{P(\text{Wash=no})} = \frac{3/6 * 2/3}{3/6} = 0.67$$

Predicted class = yes

Example of Prediction Based on the Bayes Theorem

New example: (Wash=no, y=?)

$$P(\text{Cavity=no} \mid \text{Wash=no}) = \frac{P(\text{Cavity=no})\,P(\text{Wash=no} \mid \text{Cavity=no})}{P(\text{Wash=no})} = \frac{3/6 * 1/3}{3/6} = 0.33$$

$$P(\text{Cavity=yes} \mid \text{Wash=no}) = \frac{P(\text{Cavity=yes})P(\text{Wash=no} \mid \text{Cavity=yes})}{P(\text{Wash=no})} = \frac{3/6 * 2/3}{3/6} = 0.67$$

P(Wash=no) is a normalising factor, to make the probabilities sum up to 1.
We could replace its computation by:

$$\beta = P(\text{Cavity=no})P(\text{Wash=no} \mid \text{Cavity=no}) + P(\text{Cavity=yes})P(\text{Wash=no} \mid \text{Cavity=yes})$$

Normalisation Factor α

$$\beta = \sum_{c \in Y} P(c)\, P(\mathbf{a}|c)$$

$$P(c|\mathbf{a}) = \frac{P(c)\, P(\mathbf{a}|c)}{P(\mathbf{a})} \quad \rightarrow \quad P(c|\mathbf{a}) = \frac{P(c)\, P(\mathbf{a}|c)}{\beta}$$

$$P(c|a) = \alpha\, P(c)\, P(\mathbf{a}|c) \quad \text{where} \quad \alpha = 1 / \beta$$

Bayes Theorem for d Independent Variables, where d ≥ 1

$$P(c|\mathbf{a}) = \alpha\, P(c)\, P(\mathbf{a}|c)$$

$$P(c|a_1,\ldots,a_d) = \alpha\, P(c)\, P(a_1,\ldots,a_d|c)$$

where

- P represents a probability calculated based on the frequency tables,
- c represents a class,
- a_i represents the value of independent variable x_i, $i \in \{1, 2,...,d\}$,
- d is the number of independent variables and
- α is the normalisation factor.

Example

Training Set

Perso	x_1 (Wash)	x_2 (Pain)	y (Cavity)
P1	no	yes	yes
P2	no	yes	yes
P3	yes	yes	yes
P4	yes	no	no
P5	yes	no	no
P6	no	no	no

Problem: number of possible combinations of input values becomes very large when the number of independent variables and values is large.

Model

Frequency Table	Cavity = no	Cavity = yes	Total:
Wash=no and Pain=no	1	0	1
Wash=no and Pain=yes	0	2	2
Wash=yes and Pain=no	2	0	2
Wash=yes and Pain=yes	0	1	1
Total:	3	3	6

Quiz

Training Set

Day	x_1 (Wind)	y (Play)
D1	strong	no
D2	strong	no
D3	weak	no
D4	strong	yes
D5	strong	yes

Model

Frequency Table	Play = yes	Play = no	Total:
Wind = strong	A	B	C
Wind = weak	D	E	F
Total:	G	H	I

What is the value of A,B,C,D,E,F,G,H,I in the frequency table above based on the training set provided?

And what is the value of P(Play=yes|Wind=strong)?

Further Reading

- Essential:
  - Leandro Minku's notes on "Naive Bayes — The Relationship Between The Bayes Theorem and Classification".