Naive Bayes Classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. It assumes the conditional independence of attribute values given the class:

$$p(v_1, v_2, ...v_n|c) = \Pi_i p(v_i|c)$$

Naive Bayes formula:

$$p(c|v_1, v_2, ...v_n) = p(c) \cdot \Pi_i \frac{p(c|v_i)}{p(c)}$$

Classifying a new instance $(v_1, v_2, ..., v_n)$

Let's say that our dataset has m classes $(c_1, c_2, ..., c_m)$ (target variable with m values). The Naive Bayes classifier calculates for each class c_i the conditional probability of class c_i given evidence $(v_1, v_2, ..., v_n)$

$$p(c_i|v_1, v_2, ..., v_n)$$

according to the naive Bayes formula. It classifies the example into the class with the highest probability.

Example

Will the spider catch an ant?

Past experiences of the spider catching ants:

Color	**Size**	**Time**	**Caught**
black	large	day	**YES**
white	small	night	**YES**
black	small	day	**YES**
red	large	night	**NO**
black	large	night	**NO**
white	large	night	**NO**

Ant 1: Color = white, Time = night

$$v_1 = \text{``}Color = white\text{''}$$

$$v_2 = \text{``}Time = night\text{''}$$

$$c_1 = YES$$

$$c_2 = NO$$

$$p(c_1|v_1, v_2) = \quad (1)$$

$$p(Caught = YES|Color = white, Time = night) = \quad (2)$$

$$p(Caught = YES) * \frac{p(Caught = YES|Color = white)}{p(Caught = YES)} * \frac{p(Caught = YES|Time = night)}{p(Caught = YES)} = \quad (3)$$

$$\frac{1}{2} * \frac{\frac{1}{2}}{\frac{1}{2}} * \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{4} \quad (4)$$

$$p(c_2|v_1, v_2) = \quad (5)$$

$$p(Caught = NO|Color = white, Time = night) = \quad (6)$$

$$p(Caught = NO) * \frac{p(Caught = NO|Color = white)}{p(Caught = NO)} * \frac{p(Caught = NO|Time = night)}{p(Caught = NO)} = \quad (7)$$

$$\frac{1}{2} * \frac{\frac{1}{2}}{\frac{1}{2}} * \frac{\frac{3}{4}}{\frac{1}{2}} = \frac{3}{4} \quad (8)$$

The spider will not catch the white ant at night because p(Caught=NO| Color = white, Time = night) > p(Caught=YES | Color = white, Time = night).

Ant 2: Color = black, Size = large, Time = day

$$v_1 = \text{``}Color = black\text{''}$$
$$v_2 = \text{``}Size = large\text{''}$$
$$v_3 = \text{``}Time = day\text{''}$$
$$c_1 = YES$$
$$c_2 = NO$$

$$p(c_1|v_1, v_2, v_3) = \tag{9}$$
$$p(Caught = YES|Color = black, Size = large, Time = day) = \tag{10}$$
$$p(Caught = YES) * \frac{p(Caught = YES|Color = black)}{p(Caught = YES)} * ... \tag{11}$$
$$... * \frac{p(Caught = YES|Size = large)}{p(Caught = YES)} * \frac{p(Caught = YES|Time = day)}{p(Caught = YES)} = \tag{12}$$
$$\frac{1}{2} * \frac{\frac{2}{3}}{\frac{1}{2}} * \frac{\frac{1}{4}}{\frac{1}{2}} * \frac{1}{\frac{1}{2}} = \frac{2}{3} \tag{13}$$

$$p(c_2|v_1, v_2, v_3) = \tag{14}$$
$$p(Caught = NO|Color = black, Size = large, Time = day) = \tag{15}$$
$$p(Caught = NO) * \frac{p(Caught = NO|Color = black)}{p(Caught = NO)} * ... \tag{16}$$
$$... * \frac{p(Caught = NO|Size = large)}{p(Caught = NO)} * \frac{p(Caught = NO|Time = day)}{p(Caught = NO)} * = \tag{17}$$
$$\frac{1}{2} * \frac{\frac{1}{3}}{\frac{1}{2}} * \frac{\frac{3}{4}}{\frac{1}{2}} * \frac{0}{\frac{1}{2}} = 0 \tag{18}$$

The spider will catch the large black ant at night because p(Caught=YES | Color = black, Size = large, Time = day) > p(Caught=NO | Color = black, Size = large, Time = day).

To think over:

When calculating probabilities $p(c_1|v_1, v_2)$ and $p(c_2|v_1, v_2)$ for a two class problem using naive Bayes formula, the probabilities sometimes do no sum up to 1: $p(c_2|v_1, v_2) + p(c_2|v_1, v_2) \neq 1$. Why?