Machine Learning
(Training Set and Naive Bayes Classifier)

Prof K R Chowdhary

MBM University

October 11, 2024

Discrete Attribute Vectors

⇒ In our discussions we have used single attributes to compute the Bayesian probability. However, in fact, the objects have many characteristics, i.e., many attributes, e.g., an apple is identified by all these attributes together: round shape, sweet taste, juicy material, red (or red-green) color, size of 2-inch diameters, and so on.

⇒ We cannot recognize it's class properly if only part of these attributes are made known. Attributes are represented by attribute vector, say, $\mathbf{x} = \{x_1, x_2, ..., x_n\}$.

⇒ Certain characteristics together, of an apple, may identify it as *Kashmiri apple* or *Himachal apple*. This way, an object may be in many of the one class. In domain of geometric shapes, assume two classes: *pos* and *neg*. If c_i is the label of the i^{th} class, and $\mathbf{x}$ is the vector of object we want to classify, the Bayes formula is:

$$P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i).P(c_i)}{P(\mathbf{x})}. \quad (1)$$

Discrete Attribute Vectors

Example: Training examples are given in Table 1 below. We want the machine to recognize the Class: `pos` or `neg`

Ex	Shape	Filling	Shad	Cl.
Ex1	Diam	Shading	Light	pos
Ex2	Rect	Shading	Dark	pos
Ex3	Rect	Hashing	Dark	pos
Ex4	Tria	Clear	Light	pos
Ex5	Tria	Hashing	Dark	pos
Ex6	Diam	Clear	Dark	neg
Ex7	Elli	Hashing	Dark	neg
Ex8	Elli	Clear	Light	neg
Ex9	Circ	Hashing	Dark	neg
Ex10	Circ	Shading	Light	pos

The probabilities of the individual attribute values, and of class labels are computed using the relative frequencies as discussed above.

Consider $\texttt{Shadow} = \texttt{Light}$. Hence, $P(Shadow = Light) = 4/10 = 0.4$. Also,

$$P(Dark) = \frac{6}{10} = 0.6$$

$$P(pos) = \frac{6}{10} = 0.6$$

$$P(neg) = \frac{4}{10} = 0.4$$

The conditional probability of a attribute in a class is determined by relative frequency is:

$$P(Light|pos) = \frac{3}{6} = 0.5$$
$$P(Light|neg) = \frac{1}{4} = 0.25$$
$$P(Dark|pos) = \frac{3}{6} = 0.5$$
$$P(Dark|neg) = \frac{3}{4} = 0.75$$

We calculate the conditional probabilities, using the Bayes formula:

$$P(pos|Light) = \frac{3}{4} = 0.75$$
$$P(neg|Light) = \frac{1}{4} = 0.25$$
$$P(pos|Dark) = \frac{3}{6} = 0.5$$
$$P(neg|Dark) = \frac{3}{6} = 0.5$$

First value in the above is:

$$P(pos|Light) = \frac{P(Light|pos) \times P(pos)}{P(Light)} = \frac{0.5 \times 0.6}{0.4} = 0.75$$

Computing the Vector's Probability

⇒ Note that sum of: "probability of `pos` given that it is `Light`" and "probability of `neg` given that it is `Light`" is 1. Similarly, "probability of `pos` given that it is `Dark`" and "probability of `neg` given that it is `Dark`" sums to 1.

⇒ In formula (1), denominator being same for each class, we choose the class that maximizes the numerator $P(\mathbf{x}|c_i).P(c_i)$. Here $P(c_i)$ is easy to estimate by relative frequency of c_i in the training set. However, the $P(\mathbf{x}|c_i)$ is **not** straightforward.

⇒ $P(\mathbf{x}|c_i)$ of vector $\mathbf{x}$ in (1) is the probability that a randomly selected member of class c_i is described by vector $\mathbf{x}$.

⇒ $P(\mathbf{x}|c_i)$ cannot be computed by freq., as, for class c_i we have no count for $\mathbf{x}$. In "geometric shapes" domain, size of instance space: table 1 is $5 \times 3 \times 2 = 30$., comprising different examples, but the set contains only 10. Each one represented by one training example, while the other vectors (i.e., $30 - 10 = 20$) have not been represented at all.

Computing the Vector's Probability...

⇒ Relative frequency of **x** in six positive examples in table 1 can thus be either $P(\mathbf{x}|\text{pos}) = 1/6$, when vector **x** exists in them, $P(\mathbf{x}|\text{pos}) = 0$, when vector **x** does not exist.

⇒ Any **x** identical to a training example "inherits" this example's class label. But, if $\mathbf{x} \notin$ training set, then $P(\mathbf{x}|c_i) = 0$, and $P(\mathbf{x}|c_i).P(c_i) = 0$. $\therefore$ we are unable to choose most probable class. Also, cannot calculate probability of an event that occurs only once or does not not occur (?).

⇒ Above conditions do not arise in single attributes, e.g, `Shadow`, `Shadow` = `Light` occurs three times among pos examples and once among the neg (tab-1).

⇒ Corresponding probabilities are: $P(Shadow = Light|pos)= 3/6$, $P(Shadow = Light|neg) = 1/4$. So, if an attribute consist only two /three values, chances are that each of these values is present in the training set more than once, providing better grounds for computing probabilities.

Naive Bayes Classifier:

⇒ "Naive" Bayesian classification is the optimal method of *supervised learning* if the values of the attributes of an example are independent given the class of the example.

⇒ Time required to learn a naive Bayesian classifier is linear. No learning algorithm that examines all its training data can be faster than Naive Bayes.

⇒ In Bayes, the vector **x** is labeled with class that maximizes $P(\mathbf{x}|c_i).P(c_i)$. If product's value is 0.8 for one class and 0.2 for the other, the classifier's behavior will not change even if probability estimates miss accuracy mark by ten or 20%.

Example: For attributes in table 1, apply Bayesian formula to find probabilities of vector: $\mathbf{x} = \{Shape = Rectangle, Filling = Shading, Shadow = Light\}$, if it exits in class `pos` or in `neg`. For this, we want:

$$P(pos|\mathbf{x}) = P(pos).\prod_{i=1}^{3} P(x_i|pos) \quad (2)$$

Naive Bayes Classifier...

$$P(neg|\mathbf{x}) = P(neg).\prod_{i=1}^{3} P(x_i|neg). \tag{3}$$

Procedure: Calculate numerator of Bayes formula separately for each attrib vect., then choose class with higher value. In training set, class pos has prob. $6/10$, neg has $4/10$. Probs. of attribute vector $\prod_{i=1}^{3} P(x_i|pos)$ are:

$\texttt{P(Shape} = \texttt{Rectangle}|\texttt{pos)} = 2/6$

$\texttt{P(Fillinge} = \texttt{Shading}|\texttt{pos)} = 3/6$

$\texttt{P(Shadow} = \texttt{Light}|\texttt{pos)} = 3/6$

Similarly, probabilities of attribute vector $\prod_{i=1}^{3} P(x_i|neg)$ are:

$\texttt{P(Shape} = \texttt{Rectangle}|\texttt{neg)} = 0/4$

$\texttt{P(Fillinge} = \texttt{Shading}|\texttt{neg)} = 0/4$

$\texttt{P(Shadow} = \texttt{Light}|\texttt{neg)} = 1/4$

Based on above, we obtain:

$$P(pos|\mathbf{x}) = P(pos).\prod_{i=1}^{3} P(x_i|pos)$$
$$= (0.6).\frac{2}{6}.\frac{3}{6}.\frac{3}{6}$$
$$= (0.6).\frac{18}{216} = 0.05$$

$$P(neg|\mathbf{x}) = P(neg).\prod_{i=1}^{3} P(x_i|neg)$$
$$= (0.4).\frac{0}{4}.\frac{0}{4}.\frac{1}{4}$$
$$= (0.4).\frac{0}{64} = 0.00$$

Since $P(pos|\mathbf{x}) > P(neg|\mathbf{x})$, the label of the vector **x** is class pos.