Machine Learning
(Training Set and Classifier)

Prof K R Chowdhary

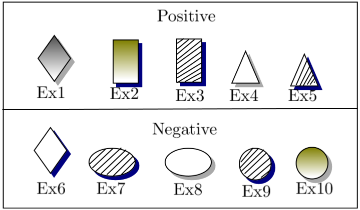
MBM University

October 04, 2024

⇒ We introduce a problem, and certain fundamental concepts that will accompany us for discussions later on.

⇒ *Pre-Classified Training Examples*: Figure shows geometrical shapes, filled with hashed-lines or grey shade or clear. They are divided into two classes: "Pos" and "neg". Task: Induce a classifier that labels future shapes as pos /neg

⇒ Task: Induce a classifier that labels future shapes as pos /neg – an algorithm capable of classifying into one of the two .



⇒ Examples: a classifier decides whether a landscape shot was taken in summer, winter, or rainy season; a Software identifies characters scribbled on a smartphone needs at least 36 classes; document categorization systems are capable of identifying hundreds, of topics.

Training Sets and Classifiers...

Table 1: The ten training examples expressed in a matrix form

Example	Shape	Filling	Shadow	Class
Ex1	Diamond	Shading	Light	Pos
Ex2	Rectangle	Shading	Dark	Pos
Ex3	Rectangle	Hashing	Dark	Pos
Ex4	Triangle	Clear	Light	Pos
Ex5	Triangle	Hashing	Dark	Pos
Ex6	Diamond	Clear	Dark	Neg
Ex7	Ellipse	Hashing	Dark	Neg
Ex8	Ellipse	Clear	Light	Neg
Ex9	Circle	Hashing	Dark	Neg
Ex10	Circle	Shading	Light	Neg

Attribute Vectors: The training examples are passed to machine as vectors. Consider five shapes: diamond, rectangle, triangle, ellipse, and circle, (table 1) make ten examples.

Training Sets and Classifiers...

A shape can be described by the conjunction, like, $(\texttt{shape} = \texttt{Diamond})\ \texttt{AND}\ (\texttt{filling} = \texttt{Shading})\ \texttt{AND}\ (\texttt{class} = \texttt{Pos})$.

Inducing A Classifier: A training set constitutes the input from which we are to induce the classifier. It can be in the form of a *Boolean function*: True for *pos* examples and False for *neg*. Expression: $[(\texttt{shape} = \texttt{Circle})\ \texttt{AND}(\texttt{filling} = \texttt{Clear})]$ is False for all negative examples. We can verify: $[(\texttt{shape} = \texttt{Triangle})]\texttt{OR}[(\texttt{shape} = \texttt{Rectangle})]$, is true always.

Brute-Force Approach: For each of 5 shapes, three filling (Shading/Hashing/Clear), the combinations are $5 \times 3 = 15$. Shade can be Light/Dark, so combinations$= 5 \times 3 \times 2 = 30$, is size of instance space. So, there are $2^{30} \approx 10^9$, alternatives.

$\Rightarrow$ In practical cases attributes are hundreds or much larger, hence checking the classes through *Boolean Satisfiability* is a complex.

The Classifier

⇒ Goal: Not to reclassify objects already classified, but classify (label) future examples. For this, we divide available pre-classified examples into two parts: 1. the *training set*, from which the classifier is induced, and 2. the *testing set*, on which the induced classifier is tested.

⇒ Fig. 1(a): training examples are used to induce a classifier,

(b): classifier is used to classify future examples.

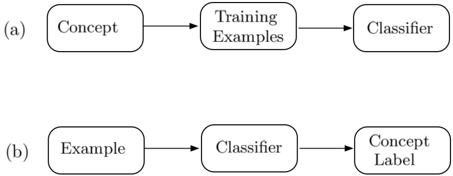


Figure 1: (a) Learning, (b) Application

Single attribute case

To understand the single attribute case, consider examples of geometric shapes in Table 2, with six positive examples (say, $N_{pos} = 6$) and four negative ($N_{neg} = 4$), and $N_{pos} + N_{neg} = N_{all}$.

Table 2: The ten training examples expressed in a matrix form

Example	Shape	Filling	Shadow	Class
Ex1	Diamond	Shading	Light	pos
Ex2	Rectangle	Shading	Dark	pos
Ex3	Rectangle	Hashing	Dark	pos
Ex4	Triangle	Clear	Light	pos
Ex5	Triangle	Hashing	Dark	pos
Ex6	Diamond	Clear	Dark	neg
Ex7	Ellipse	Hashing	Dark	neg
Ex8	Ellipse	Clear	Light	neg
Ex9	Circle	Hashing	Dark	neg
Ex10	Circle	Shading	Light	pos

Let us assume that these ten examples have correct representation of the data, the probability that a randomly selected shape is positive is:

$$P(pos) = \frac{N_{pos}}{N_{all}} = \frac{6}{10} = 0.6 \tag{1}$$

Considering the same examples as mentioned above, we reorganize them as shown in Fig. 2, such that the set of Positive and negative remains the same and there is a new classification *Round*, which comprise circles and ellipses, and it has finite intersection with the set Dark, and all the objects are in the set "All".

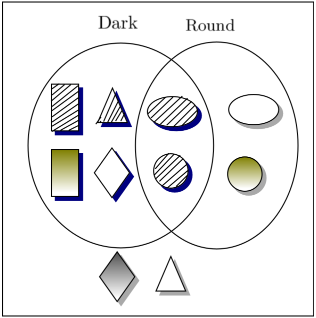


Figure 2: Computing Bayes probabilities using geometric shapes

Computing Conditional probabilities

For this, we take one attribute, say, `Shadow`. Four examples are with `Shadow` = `Light`, in that 3 are *pos* ($N_{pos|Light} = 3$), one is ($N_{neg|Light} = 1$). So, the "conditional probability of an example being positive, given that it is Light" is 75% and also, the condition probability of *neg* given that it is Light is 25%:

$$P(pos|Light) = \frac{N_{pos|Light}}{N_{Light}} = \frac{3}{4} \quad (2)$$

$$P(neg|Light) = \frac{N_{neg|Light}}{N_{Light}} = \frac{1}{4} \quad (3)$$

And, 6 examples with `Shadow` = `Dark`. From Fig. 2, *prior probabilities* are:

$$P(Dark) = \frac{6}{10} = 0.60,$$
$$P(Round) = \frac{4}{10} = 0.40$$

and *conditional probabilities*:

$$P(Dark|Round) = \frac{2}{4} = 0.50,$$
$$P(Round|Dark) = \frac{2}{6} = 0.33,$$
$$P(Dark, Round) = \frac{2}{10} = 0.20.$$

Computing Conditional Probabilities...

From the equations (2) and (3), having given that the $\texttt{Shade} = \texttt{Light}$, the positive and negative examples have probabilities of 0.75 and 0.25, respectively. So, they do into different classes.

$$P(Shading|pos) = \frac{3}{6} = 0.50, \quad (4)$$

The probability of `Shading` is, $3/10 = 0.30$, and the probability of `Clear`, given that it is positive is,

$$P(Clear|pos) = \frac{1}{6} = 0.16, \quad (5)$$

while probability of `Cleaar` is $3/10 = 0.30$. Note that probabilities of Shading and Clear are, 0.50, 0.16. So, $P(Shading|pos) > P(Shading)$, as $0.50 > 0.30$. And, $P(Clear|pos) < P(Clear)$, as $0.16 < 0.30$.