Naive Bayes Classifier example

Eric Meisner

November 22, 2003

1 The Classifier

The Bayes Naive classifier selects the most likely classification V_{nb} given the attribute values $a_1, a_2, \ldots a_n$. This results in:

$$V_{nb} = \text{argmax}_{v_j \in V} P(v_j) \prod P(a_i|v_j) \tag{1}$$

We generally estimate $P(a_i|v_j)$ using m-estimates:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m} \tag{2}$$

where:

n =	the number of training examples for which $v = v_j$
n_c =	number of examples for which $v = v_j$ and $a = a_i$
p =	a priori estimate for $P(a_i|v_j)$
m =	the equivalent sample size

2 Car theft Example

Attributes are Color , Type , Origin, and the subject, stolen can be either yes or no.

2.1 data set

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

2.2 Training example

We want to classify a Red Domestic SUV. Note there is no example of a Red Domestic SUV in our data set. Looking back at equation (2) we can see how to compute this. We need to calculate the probabilities

```
P(Red|Yes), P(SUV|Yes), P(Domestic|Yes) ,

P(Red|No) , P(SUV|No), and P(Domestic|No)
```

and multiply them by P(Yes) and P(No) respectively . We can estimate these values using equation (3).

```
Yes:                    No:
    Red:                    Red:
        n = 5                   n = 5
        n_c= 3                  n_c = 2
        p = .5                  p = .5
        m = 3                   m = 3
    SUV:                    SUV:
        n = 5                   n = 5
        n_c = 1                 n_c = 3
        p = .5                  p = .5
        m = 3                   m = 3
    Domestic:               Domestic:
        n = 5                   n = 5
        n_c = 2                 n_c = 3
        p = .5                  p = .5
        m = 3                   m =3
```

Looking at $P(Red|Yes)$, we have 5 cases where v_j = Yes , and in 3 of those cases a_i = Red. So for $P(Red|Yes)$, n = 5 and n_c = 3. Note that all attribute are binary (two possible values). We are assuming no other information so, p = 1 / (number-of-attribute-values) = 0.5 for all of our attributes. Our m value is arbitrary, (We will use m = 3) but consistent for all attributes. Now we simply apply eqauation (3) using the precomputed values of n , n_c, p, and m.

$$P(Red|Yes) = \frac{3 + 3 * .5}{5 + 3} = .56 \qquad P(Red|No) = \frac{2 + 3 * .5}{5 + 3} = .43$$
$$P(SUV|Yes) = \frac{1 + 3 * .5}{5 + 3} = .31 \qquad P(SUV|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$
$$P(Domestic|Yes) = \frac{2 + 3 * .5}{5 + 3} = .43 \qquad P(Domestic|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$

We have $P(Yes) = .5$ and $P(No) = .5$, so we can apply equation (2). For $v = Yes$, we have

```
P(Yes) * P(Red | Yes) * P(SUV | Yes) * P(Domestic|Yes)

    =  .5 * .56 * .31 * .43   = .037
```

and for $v = No$, we have

```
P(No) * P(Red | No) * P(SUV | No) * P (Domestic | No)

    = .5 * .43 * .56 * .56 = .069
```

Since $0.069 > 0.037$, our example gets classified as 'NO'