Naive Bayesian Classifiers

We have a database describing customers of our company.
Now we are assessing a new customers and do not know all his/her attribute values.
We want to use the Bayes' Rule to predict values of his/her still unknown attributes.
We are implicitly assuming he/she is extracted form the same population represented in the database.
Our prediction will be made by analogy with already known customers.

Let us take an attribute, say *Sex*. We want to "predict" the customer's gender. This is a *classification problem*.
In general, we have a collection of classes C_1, ..., C_n and we want to put the new customer in one of them. In this case, we have only two classes, Males and Females.
We make a collection of hypotheses H_1, ..., H_n, each of them in correspondence with a class. Hypothesis H_i means "the new customer's target attribute value belongs to class C_i". In the example we have 2 hypotheses, belongs-to-males and belongs-to-females.

The *Bayes' Classifier* works as follows.
We have a certain evidence E which is the content of the training dataset.
We compute the posterior $P(H_i \mid E)$ using Bayes' Rule for each hypothesis H_i.
We choose the H^* hypothesis which maximizes the posterior.
That is the winner in the competition among hypothesis.

Implementing this method in software code is simple in principle, but it can offer some problems in practice, because sometime it can require a lot of computation and/or a lot of data.

Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex
Yes	No	No	No	Male
Yes	Yes	Yes	Yes	Female
No	No	No	No	Male
Yes	Yes	Yes	Yes	Male
Yes	No	Yes	No	Female
No	No	No	No	Female
Yes	Yes	Yes	Yes	Male
No	No	No	No	Male
Yes	No	No	No	Male
Yes	Yes	Yes	No	Female

The database describes 10 customers.
All variables are categorical. It is not necessary, only for simplicity.
We can choose any variable as answer to be predicted.
Normally it is likely to be *Life Insurance Promotion*, in this example it will be *Sex*.
A new customer has attributes

Magazine Promotion = *Yes* Watch Promotion = *Yes*

Life Insurance Promotion = *No* Credit Card Insurance = *No*

Sex = ?

We use the Bayes' Rule to evaluate probability of two hypothesis:

1. the customer is male
2. the customer is female

For convenience we build a table representing the distribution of the output variable as function of input variables.

	Magazine Promotion		**Watch Promotion**		**Life Insurance Promotion**		**Credit Card Insurance**	
Sex	Male	Female	Male	Female	Male	Female	Male	Female
Yes	4	3	2	2	2	3	2	1
No	2	1	4	2	4	1	4	3
Ratio: yes/ total	4/ 6	3/ 4	2/ 6	2/ 4	2/ 6	3/ 4	2/ 6	1/ 4
Ratio: no/ total	2/ 6	1/ 4	4/ 6	2/ 4	4/ 6	1/ 4	4/ 6	3/ 4

Now we compute probability of teh new customer being a male.
Hypothesis *H* is *Sex = Male*
Evidence *E* is *Magazine = Yes and Watch = Yes and Life = No and Credit = No*

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

$$P(sex = male \mid E) = \frac{P(E \mid sex = male)\, P(sex = male)}{P(E)}$$

	Magazine Promotion		Watch Promotion		Life Insurance Promotion		Credit Card Insurance	
Sex	Male	Female	Male	Female	Male	Female	Male	Female
Yes	4	3	2	2	2	3	2	1
No	2	1	4	2	4	1	4	3
Ratio: yes/ total	4/ 6	3/ 4	2/ 6	2/ 4	2/ 6	3/ 4	2/ 6	1/ 4
Ratio: no/ total	2/ 6	1/ 4	4/ 6	2/ 4	4/ 6	1/ 4	4/ 6	3/ 4

$$P(sex = male \mid E) = \frac{P(E \mid sex = male)\, P(sex = male)}{P(E)}$$

The problematic point is the likelihood con P(E | Sex = Male).
Evidence E is the conjunction of 4 evidence items.
Now we state a strong hypothesis: these 4 evidence items are independent.
For a male, probability of buying one of the 4 promotions is independent on the other 3.
Intuitively, it is not true. Why such an hypothesis?
The reason is that for independent events this useful property holds:

$$P(X_1,..., X_n \mid Y = y) = \prod_{i=1}^{n} P(X_i \mid Y = y)$$

$$P(X_1,\ldots,X_n \mid Y=y) = \prod_{i=1}^{n} P(X_i \mid Y=y)$$

Here the assumption here is that all conditional $X_i \mid y$ are independent: knowledge about one of them is not useful to estimate the others.
Under this assumption, we can compute the likelihood as product of 4 likelihoods, each one simpler than the original.
One advantage can be in terms of computation, though not in simple cases like here.
Another is that it is possible to be not able to compute the left side because you do not have examples of that kind in the database.
If you have 30 variables X_i each one with only 2 possible values, combinations are one billion. Probably you do not have so many examples of customers in the database, and even if you have them, each sample is not statistically meaningful.
Sometimes we lack data necessary to compute the original form of likelihood in member side.
But it is likely we have samples enough for each one-variable likelihood.
So, it would be extremely useful to use the right side.

$$P(X_1,\ldots,X_n \mid Y=y)=\prod_{i=1}^{n} P(X_i \mid Y=y)$$

We do not claim that assumption of independence is true.
We only hope it is not-so-false to invalid our conclusions. If the variables are not-too-dependent, our assumption causes an acceptable approximation in computation, while enabling us to use much larger samples.
In practice, often this hypothesis does not damage conclusions too much and give use a more practical formulation of Bayes' Rule:

$$P(H \mid E)=\frac{P(H)\bullet\prod_{i=1}^{n} P(X_i \mid H)}{P(E)}$$

A Bayes' Classifier using this formula is named *Naïve Bayesian Classifier*.
It is naïve because it exploits an hypothesis very unlikely to be true.

	Magazine Promotion		Watch Promotion		Life Insurance Promotion		Credit Card Insurance	
Sex	Male	Female	Male	Female	Male	Female	Male	Female
Yes	4	3	2	2	2	3	2	1
No	2	1	4	2	4	1	4	3
Ratio: yes/ total	4/ 6	3/ 4	2/ 6	2/ 4	2/ 6	3/ 4	2/ 6	1/ 4
Ratio: no/ total	2/ 6	1/ 4	4/ 6	2/ 4	4/ 6	1/ 4	4/ 6	3/ 4

$$P(sex = male \mid E) = \frac{P(E \mid sex = male)\, P(sex = male)}{P(E)}$$

Now we compute likelihood as product of 4 simpler likelihoods, each one simply estimated with its empirical frequency in the database.

P(Magazine = Yes | Sex = Male) = 4 / 6
P(Watch = Yes | Sex = Male) = 2 / 6
P(Life = No | Sex = Male) = 4 / 6
P(Credit = No | Sex = Male) = 4 / 6
*P(E | Sex = Male) = 4/6 * 2/6 * 4/6 * 4/6 = 8/81*

For a male customer, probability of evidence being just the visible one is 8/81, about 10%.

This does not mean that probability of the customer being a female is 90%. Tis number is likelihood, just one out of 3 factors.

Now we have to compute prior and marginal, in order to finally compute the posterior.

$$P(sex = male \mid E) = \frac{P(E \mid sex = male)\, P(sex = male)}{P(E)}$$

	Magazine Promotion		Watch Promotion		Life Insurance Promotion		Credit Card Insurance	
Sex	Male	Female	Male	Female	Male	Female	Male	Female
Yes	4	3	2	2	2	3	2	1
No	2	1	4	2	4	1	4	3
Ratio: yes/ total	4/ 6	3/ 4	2/ 6	2/ 4	2/ 6	3/ 4	2/ 6	1/ 4
Ratio: no/ total	2/ 6	1/ 4	4/ 6	2/ 4	4/ 6	1/ 4	4/ 6	3/ 4

$$P(sex = male \mid E) = \frac{P(E \mid sex = male)\, P(sex = male)}{P(E)}$$

Now we compute *P(Sex = Male).*
This is the prior, the probability of the customer being a male when we have no data about him/her.
It equals 3/5, because we have 6 males out of 10 customers.
P(Sex = Male) = 3/5.

Now we know the upper side of the Bayes' Rule:
*P(E | Sex = Male) P(Sex = Male) = 8/81 * 3/5 = 0,0593.*

	Magazine Promotion		Watch Promotion		Life Insurance Promotion		Credit Card Insurance	
Sex	Male	Female	Male	Female	Male	Female	Male	Female
Yes	4	3	2	2	2	3	2	1
No	2	1	4	2	4	1	4	3
Ratio: yes/ total	4/ 6	3/ 4	2/ 6	2/ 4	2/ 6	3/ 4	2/ 6	1/ 4
Ratio: no/ total	2/ 6	1/ 4	4/ 6	2/ 4	4/ 6	1/ 4	4/ 6	3/ 4

$$P(sex = male \mid E) = \frac{P(E \mid sex = male)\, P(sex = male)}{P(E)}$$

Now we have to estimate probability of observed data being just E:

Magazine Promotion = Yes, Watch Promotion = Yes

Life Insurance Promotion = No, Credit Card Insurance = No

Really, this is not necessary.

The competitor hypothesis posterior, *Sex = Female*, will be computed exactly in the same way as a ratio with another upper side and the same lower side.

In order to choose the maximum posterior, it is not necessary to compute the whole ratio, only the upper side.

	Magazine Promotion		Watch Promotion		Life Insurance Promotion		Credit Card Insurance	
Sex	Male	Female	Male	Female	Male	Female	Male	Female
Yes	4	3	2	2	2	3	2	1
No	2	1	4	2	4	1	4	3
Ratio: yes/ total	4/ 6	3/ 4	2/ 6	2/ 4	2/ 6	3/ 4	2/ 6	1/ 4
Ratio: no/ total	2/ 6	1/ 4	4/ 6	2/ 4	4/ 6	1/ 4	4/ 6	3/ 4

$$P(sex = male \mid E) = \frac{P(E \mid sex = male)\, P(sex = male)}{P(E)}$$

Let us move to assess the hypothesis *Sex = Female*
First, *P(E | Sex = Female).*

P(Magazine = Yes | Sex = Female) = 3 / 4
P(Watch = Yes | Sex = Female) = 2 / 4
P(Life = No | Sex = Female) = 1 / 4
P(Credit = No | Sex = Female) = 3 / 4
*P(E | Sex = Female) = 3/4 * 2/4 * 1/4 * 2/4 = 9/128*
This likelihood is less than for males, which was 8/81.
But we have to remember that females are fewer than males in the database (4 versus 6).

	Magazine Promotion		Watch Promotion		Life Insurance Promotion		Credit Card Insurance	
Sex	Male	Female	Male	Female	Male	Female	Male	Female
Yes	4	3	2	2	2	3	2	1
No	2	1	4	2	4	1	4	3
Ratio: yes/ total	4/ 6	3/ 4	2/ 6	2/ 4	2/ 6	3/ 4	2/ 6	1/ 4
Ratio: no/ total	2/ 6	1/ 4	4/ 6	2/ 4	4/ 6	1/ 4	4/ 6	3/ 4

$$P(sex = male \mid E) = \frac{P(E \mid sex = male)\, P(sex = male)}{P(E)}$$

Now, the prior *P(Sex = Female).*
It is 2/5, because there are 4 females out of 10 males.
P(Sex = Female) = 2/5.
The upper side of ratio is:
*P(E | Sex = Female) P(Sex = Female) = 9/128 * 2/5 = 0,0281.*

Now we could compute the marginal P(E):

Magazine Promotion = YesWatch Promotion = Yes

Life Insurance Promotion = No Credit Card Insurance = No

But it is not necessary (see above).

We estimated:

$$P(Sex = Male \mid E) = 0{,}0593 / P(E)$$
$$P(Sex = Female \mid E) = 0{,}0281 / P(E)$$

The winning hypothesis is that the unknown gender of the new customer is male, considering his/her behavior.

Naïve Bayesian Classifier are often surprisingly effective, notwithstanding the assumption of independence, which is often simplistic.
An explanation of such effectiveness is that we are using the hypothesis to estimate only a *ranking*: the most likely hypothesis, the second most probable and so on. We are not searching for a *scoring*, sayng us how much this hypothesis is more likely than that.
So, until the approximation due to the independence assumption causes small distortion in the scoring, it is likely the “winner” will remain the true one, or maybe the second or third most probable one. In this case, the true ranking is only affected in a limited measure in comparison, letting our classifications and prediction good enough..