This method is known as an optimal classifier because it provides the best possible classification system. Another classification system, given the same data, can only hope to classify unseen data as well as this method—it cannot do better than the optimal classifier, on average.

12.9 The Naïve Bayes Classifier

The **naïve Bayes classifier** is a simple but effective learning system. Each piece of data that is to be classified consists of a set of attributes, each of which can take on a number of possible values. The data are then classified into a single classification.

To identify the best classification for a particular instance of data ($d_1, \ldots, d_n$), the posterior probability of each possible classification is calculated:

$$P(c_i | d_1, \ldots, d_n)$$

where c_i is the ith classification, from a set of $|c|$ classifications.

The classification whose posterior probability is highest is chosen as the correct classification for this set of data. The hypothesis that has the highest posterior probability is often known as the **maximum a posteriori**, or MAP hypothesis. In this case, we are looking for the MAP classification.

To calculate the posterior probability, we can use Bayes' theorem and rewrite it as

$$\frac{P(d_1, \ldots, d_n | c_i) \cdot P(c_i)}{P(d_1, \ldots, d_n)}$$

Because we are simply trying to find the highest probability, and because $P(d_1, \ldots, d_n)$ is a constant independent of c_i, we can eliminate it and simply aim to find the classification c_i, for which the following is maximized:

$$P(d_1, \ldots, d_n | c_i) \cdot P(c_i)$$

The naïve Bayes classifier now assumes that each of the attributes in the data item is independent of the others, in which case $P(d_1, \ldots, d_n | c_i)$ can be rewritten and the following value obtained:

$$P(c_i) \cdot \prod_{j=1}^{n} P(d_j | c_i)$$

The naïve Bayes classifier selects a classification for a data set by finding the classification c_i for which the above calculation is a maximum.

For example, let us suppose that each data item consists of the attributes x, y, and z, where x, y, and z are each integers in the range 1 to 4.

The available classifications are A, B, and C.

The example training data are as follows:

x	y	z	Classification
2	3	2	A
4	1	4	B
1	3	2	A
2	4	3	A
4	2	4	B
2	1	3	C
1	2	4	A
2	3	3	B
2	2	4	A
3	3	3	C
3	2	1	A
1	2	1	B
2	1	4	A
4	3	4	C
2	2	4	A

Hence, we have 15 pieces of training data, each of which has been classified. Eight of the training data are classified as A, four as B, and three as C.

Now let us suppose that we are presented with a new piece of data, which is

$$(x = 2, y = 3, z = 4)$$

We need to obtain the posterior probability of each of the three classifications, given this piece of training data. Note that if we were to attempt to calculate $P(c_i|x = 2, y = 3, z = 4)$ without having made the simplifying step that

we took above, in assuming that the attribute values are independent of each other, then we would need to have had many more items of training data to proceed. The naïve Bayes classifier requires far fewer items of training data.

We must now calculate each of the following:

$$P(A) \cdot P(x = 2|A) \cdot P(y = 3|A) \cdot P(z = 4|A)$$
$$P(B) \cdot P(x = 2|B) \cdot P(y = 3|B) \cdot P(z = 4|B)$$
$$P(C) \cdot P(x = 2|C) \cdot P(y = 3|C) \cdot P(z = 4|C)$$

Hence, for classification A, we obtain the following:

$$\frac{8}{15} \cdot \frac{5}{8} \cdot \frac{2}{8} \cdot \frac{4}{8} = 0.0417$$

This was calculated by observing that of the 15 items of training data, 8 were classified as A, and so $P(A) = 8/15$. Similarly, of the eight items of training data that were classified as A, five had $x = 2$, two had $y = 3$, and four had $z = 4$, and so $P(x = 2|A) = 5/8$, $P(y = 3|A) = 2/8$, and $P(z = 4|A) = 4/8$.

Similarly, we obtain the posterior probability for category B:

$$\frac{4}{15} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{2}{4} = 0.0083$$

and for category C:

$$\frac{3}{15} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} = 0.015$$

Hence, category A is chosen as the best category for this new piece of data, with category C as the second best choice.

Let us now suppose that we are to classify the following piece of unseen data:

$$(x = 1, y = 2, z = 2)$$

As before, we would calculate the posterior probability for A. However, in calculating the probabilities for B and C, we would have problems. In the case of category B, we would have

$$P(x = 1|B) = 1/5$$
$$P(y = 2|B) = 1/5$$
$$P(z = 2|B) = 0$$

Because there are no training examples with $z = 2$ that were classified as B, we have a posterior probability of 0. Similarly, for category C, we end up with

$$P(x = 1|C) = 0$$
$$P(y = 2|C) = 0$$
$$P(z = 2|C) = 0$$

In this case, we clearly must select category *A* as the best choice for the data, but it appears to be based on a fairly inadequate comparison because insufficient training data were available to properly compute posterior probabilities for the other categories.

This problem can be avoided by using the **m-estimate**, as follows:

We wish to determine the probability of a particular attribute value, given a particular classification, such as $P(x = 1|C)$. We will estimate this probability according to the following formula:

$$\frac{a + mp}{b + m}$$

where a = the number of training examples that exactly match our requirements (e.g., for $P(x = 1|C)$, a is the number of training examples where $x = 1$ and that have been categorized as C. In this example, a is 0); b = the number of training examples that were classified in the current classification (i.e., for $P(x = 1|C)$, b is the number of items of training data that were given classification C); p = an estimate of the probability that we are trying to obtain (usually this is obtained by simply assuming that each possible value is equally likely—hence, in our example, for $P(x = 1|C)$, $p = 1/4 = 0.25$, as it would be for each of the other three possible values for x); m is a constant value, known as the **equivalent sample size.**

For example, let us use an equivalent sample size of 5 and determine the best classification for $(x = 1, y = 2, z = 2)$:

For category *A*, we first need to calculate the probability for each of the three attributes.

Hence, for $x = 1$:

$$\frac{2 + 5/4}{8 + 5} = 0.25$$

For $y = 2$:

$$\frac{3+5/4}{8+5} = 0.33$$

For $z = 2$:

$$\frac{1+5/4}{8+5} = 0.17$$

Hence, the posterior probability estimate for A is

$$\frac{8}{15} \cdot 0.25 \cdot 0.33 \cdot 0.17 = 0.0076$$

Similarly, we can now obtain posterior probability estimates for categories B and C:

For category B, we obtain the following three probabilities:

$$\frac{1+5/4}{5+5} = 0.225, \quad \frac{2+5/4}{5+5} = 0.325, \quad \frac{0+5/4}{5+5} = 0.125$$

This gives us a posterior probability for category B as follows:

$$\frac{5}{15} \cdot 0.225 \cdot 0.325 \cdot 0.125 = 0.0091$$

Finally, the posterior probability for category C can be obtained. We note first that each of the three probabilities is the same because none of the attribute values occur in the training data with category C. Hence, the probability we use will be

$$\frac{0+5/4}{3+5} = 0.156$$

Hence, the posterior probability for category C is as follows:

$$\frac{3}{15} \cdot 0.156 \cdot 0.156 \cdot 0.156 = 0.0008$$

Hence, using this estimate for probability, we find that category B is the best match for the new data, and not category A as would have been obtained using the simpler probability estimates.

It is possible to further simplify the naïve Bayes classifier by considering the values to be positionless within each item of data. In other words, when considering a new item of data, rather than assigning values to three attributes, we can simply think of the data as consisting of three values, whose order is arbitrary.

For example, consider the piece of new data (2, 3, 4).

In this case, we use the same method as before, but rather than considering the probability that, for example, $x = 2$ when an item is classified as A, we simply consider the probability that any attribute has value 2.

This simplified version of the naïve Bayes classifier is often used in text classification applications. Here, the categories are often simply "relevant" and "irrelevant," and the data to be classified consist of the words contained within textual documents. For example, an item of data might be ("the," "cat," "sat," "on," "the," "mat"). Training data would be presented in the form of a set of documents that has been preclassified as relevant and a set that has been preclassified as irrelevant. This form of textual analysis is discussed in more detail in Chapter 20, which is concerned with information retrieval and natural language processing.

12.10 Collaborative Filtering

A further practical use for Bayesian reasoning is in **collaborative filtering**. Collaborative filtering is a technique that is increasingly used by online stores (such as Amazon.com) to provide plausible suggestions to customers based on their previous purchases. The idea behind collaborative filtering can be stated very simply: if we know that Anne and Bob both like items A, B, and C, and that Anne likes D, then it is reasonable to suppose that Bob would also like D.

Collaborative filtering can be implemented in a number of ways, and the Bayesian inference has proved to be a successful method. This involves working with posterior probabilities such as the following:

$$P(\text{Bob Likes Z} \mid \text{Bob likes A, Bob likes B, . . ., Bob Likes Y})$$

Clearly, for this mechanism to work accurately, large amounts of data must be collected. Information about thousands of individuals is needed, and information is required about dozens or hundreds of items for each individual. In the case of commerce sites, this information can be collected on