55 NAÏVE BAYES CLASSIFIER

The optimal Bayes classifier (52.8) requires knowledge of the conditional probability distribution $\mathbb{P}(\boldsymbol{r}=r|\boldsymbol{h}=h)$, which is generally unavailable. In this and the next few chapters, we describe data-based *generative* methods that approximate the joint probability distribution $f_{\boldsymbol{r},\boldsymbol{h}}(r,h)$, or its components $\mathbb{P}(\boldsymbol{r}=r)$ and $f_{\boldsymbol{h}|\boldsymbol{r}}(h|\boldsymbol{r})$, directly from the data. Once these components are estimated, they can then used to learn the desired probabilities $\mathbb{P}(\boldsymbol{r}=r|\boldsymbol{h}=h)$ by means of Bayes rule and to perform classification. Among these methods we list the *naïve Bayes classifier* of this chapter, the linear and Fisher discriminant analysis (LDA, FDA) methods of the next chapter, and the logistic regression method of Chapter 59.

The naïve classifier is a sub-optimal construction that relies on a certain *independence* assumption. Although the assumption rarely holds in practice, the resulting classifier has become popular and leads to competitive performance in many applications involving text segmentation, document classification, spam filtering, or medical diagnosis. The naïve Bayes classifier is an example of a *supervised* learning procedure because its training requires access to a collection of feature vectors and their respective labels. The training data is used to estimate the priors $\mathbb{P}(\boldsymbol{r}=r)$ and to fit Bernoulli or multinomial distributions to model the conditional $f_{\boldsymbol{h}|\boldsymbol{r}}(h|r)$.

55.1 INDEPENDENCE CONDITION

We start by describing the independence assumption that will facilitate the evaluation of the Bayes classifier and lead to its naïve implementation. Specifically, we will assume that:

(a) (**Discrete attributes**) The individual entries (or attributes) of the feature vector $\boldsymbol{h}\in\mathbb{R}^M$, denoted by $\{\boldsymbol{h}(1),\boldsymbol{h}(2),\dots,\boldsymbol{h}(M)\}$, assume *discrete* values (i.e., they are not continuous random variables). Later, in Secs. 55.4 and 56.2, we will consider the situation in which the entries of the feature vector are *continuously*-distributed.

(b) (**Conditionally independent attributes**) The individual entries $\{\boldsymbol{h}(m)\}$ are conditionally independent of each other given the class variable $\boldsymbol{r}$, so that the joint probability of any two entries decouples into the product of the individual probabilities:

$$\begin{aligned}&\mathbb{P}\Big(\boldsymbol{h}(k)=a,\boldsymbol{h}(\ell)=b|\boldsymbol{r}=r\Big)\\&=\mathbb{P}\Big(\boldsymbol{h}(k)=a|\boldsymbol{r}=r\Big)\;\times\;\mathbb{P}\Big(\boldsymbol{h}(\ell)=b|\boldsymbol{r}=r\Big)\end{aligned}\tag{55.1}$$

for any $k \neq \ell$.

Let π_r represent the prior probability for each class $\boldsymbol{r}=r$, namely,

$$\pi_r \;\overset{\Delta}{=}\; \mathbb{P}(\boldsymbol{r}=r),\quad r=1,2,\ldots,R \tag{55.2}$$

Now, given a feature vector h, we would like to determine its most likely label according to the Bayes classifier construction. Using Bayes rule (3.42c) for discrete random variables, we can express the desired conditional probability in the form:

$$\mathbb{P}(\boldsymbol{r}=r|\boldsymbol{h}=h)=\frac{\mathbb{P}(\boldsymbol{r}=r)\,\mathbb{P}(\boldsymbol{h}=h|\boldsymbol{r}=r)}{\mathbb{P}(\boldsymbol{h}=h)} \tag{55.3}$$

Since the quantity in the denominator, $\mathbb{P}(\boldsymbol{h}=h)$, is independent of $\boldsymbol{r}$, we can ignore its presence and note that in order to maximize $\mathbb{P}(\boldsymbol{r}=r|\boldsymbol{h}=h)$ over r it is sufficient to maximize the numerator so that the label for h can be found by solving (where we are using the bullet superscript notation to refer to this optimal construction):

$$r^{\bullet}(h)\;\overset{\Delta}{=}\;\underset{1\leq r\leq R}{\text{argmax}}\;\Big\{\pi_r\,\mathbb{P}(\boldsymbol{h}=h|\boldsymbol{r}=r)\Big\} \tag{55.4}$$

We therefore transformed the problem of determining the label for h into one that requires evaluation of the *reverse* conditional probability $\mathbb{P}(\boldsymbol{h}=h|\boldsymbol{r}=r)$. It is at this stage that the independence assumption becomes useful. This is because it allows us to write the factorization:

$$\mathbb{P}(\boldsymbol{h}=h|\boldsymbol{r}=r)\;=\;\prod_{m=1}^{M}\mathbb{P}\Big(\boldsymbol{h}(m)=h(m)|\boldsymbol{r}=r\Big) \tag{55.5}$$

Substituting into (55.4) and transforming the right-hand side into the logarithmic scale to avoid working with small numbers, we arrive at:

(**Bayes classifier under independence assumption**) (55.6)

$$r^{\bullet}(h)\;\overset{\Delta}{=}\;\underset{1\leq r\leq R}{\text{argmax}}\;\left\{\log(\pi_r)\;+\;\sum_{m=1}^{M}\log\Big(\mathbb{P}(\boldsymbol{h}(m)=h(m)|\boldsymbol{r}=r)\Big)\right\}$$

Example 55.1 (**Document classification**) Consider an application in which we are interested in classifying a newspaper document into one of four classes defined as follows:

$$\left\{\begin{array}{lcl}\boldsymbol{r}=1 & \longrightarrow & \text{article discusses sports}\\ \boldsymbol{r}=2 & \longrightarrow & \text{article discusses politics}\\ \boldsymbol{r}=3 & \longrightarrow & \text{article discusses movies}\end{array}\right. \tag{55.7}$$

Assume further, for this contrived example, that we extract four attributes from each document and collect them into a 4−dimensional feature vector, $\boldsymbol{h} \in \mathbb{R}^4$, where each entry of $\boldsymbol{h}$ counts the total number of times that the words below appear in the article:

$$\boldsymbol{h}(1): \{\text{football, basketball, baseball}\} \tag{55.8a}$$
$$\boldsymbol{h}(2): \{\text{President, Congress, election}\} \tag{55.8b}$$
$$\boldsymbol{h}(3): \{\text{actor, actress, theater}\} \tag{55.8c}$$
$$\boldsymbol{h}(4): \{\text{inflation, market, consumer}\} \tag{55.8d}$$

In this case, we have $R = 3$ classes and $M = 4$ attributes. Obviously, in actual text classification systems, the construction of the feature space is more comprehensive than shown here and will take into account several other aspects of the document.

Given $\boldsymbol{r} = 2$ (the document discusses politics), the independence assumption amounts to saying that the number of times that the words {President, Congress, election} appear in the document is, for example, conditionally independent of the number of times that the words {football, basketbal, baseball} appear in the same document.

55.2 MODELING THE CONDITIONAL DISTRIBUTION

Determination of the Bayes classifier by means of (55.6) still requires knowledge of the *reverse* conditional probability $\mathbb{P}(\boldsymbol{h} = h|\boldsymbol{r} = r)$. Since we are assuming $\boldsymbol{h}$ to be discrete, we can consider two distributions that are particularly useful to model such probabilities.

Bernoulli distribution

In one model, we assume each attribute $\boldsymbol{h}(m) \in \{0, 1\}$ follows a Bernoulli distribution and is binary-valued. Situations like this arise, for example, when $h(m)$ is declaring the presence of a certain attribute or not (such as whether an object is hot or cold, blue or yellow, and so forth). Let p_{rm} denote the success probability, i.e., the likelihood that $\boldsymbol{h}(m)$ assumes the value one under class $\boldsymbol{r} = r$:

$$p_{rm} \;\stackrel{\Delta}{=}\; \mathbb{P}\Big(h(m) = 1|\boldsymbol{r} = r\Big) \tag{55.9}$$

Note that we are attaching two subscripts to p_{rm}: the subscript r indicates that the value of p_{rm} depends on the class variable, and the subscript m is the index of the attribute. Thus, the value of p_{rm} is referring to the likelihood that the $m-$th attribute is active given that the feature vector belongs to class r. For this same attribute, but under another class r', the value $p_{r'm}$ can be different. Using (55.9), we can write

$$\mathbb{P}\Big(\boldsymbol{h}(m) = h(m)|\boldsymbol{r} = r\Big) \;=\; p_{rm}^{h(m)}\,(1 - p_{rm})^{1-h(m)}, \quad h(m) \in \{0, 1\} \tag{55.10}$$

In this way, we can determine the probabilities $\mathbb{P}(\boldsymbol{h} = h|\boldsymbol{r} = r)$ from knowledge of the $\{p_{rm}\}$.

Multinomial distribution

More generally, we allow each attribute in the feature vector to assume a multitude of discrete levels (e.g., red, blue, green), and we let $h(m)$ measure the number of times that attribute m has been observed (e.g., how many times the color red occurred, the color blue, and the color green). In this case, the variables $\{\boldsymbol{h}(1), \boldsymbol{h}(2), \ldots, \boldsymbol{h}(M)\}$ follow a multinomial distribution. Let p_{rm} denote the likelihood of observing attribute m under class r. These probabilities satisfy

$$\sum_{m=1}^{M} p_{rm} = 1, \;\; \forall\, r \in \{1, 2, \ldots, R\} \tag{55.11}$$

and, using expression (5.34), we have

$$\mathbb{P}(\boldsymbol{h} = h | \boldsymbol{r} = r) \;=\; \frac{\Big(\sum_{m=1}^{M} h(m)\Big)!}{h(1)!\, h(2)! \ldots\, h(M)!}\, p_{r1}^{h(1)}\, p_{r2}^{h(2)} \;\ldots\; p_{rM}^{h(M)} \tag{55.12}$$

Observe that this expression provides the conditional probability of $\boldsymbol{h}$ directly rather than of its individual entries, as was the case with (55.10). This is of course sufficient for use in (55.4).

55.3 ESTIMATING THE PRIORS

We are now ready to derive the naïve Bayes classifier. One of the main difficulties in implementing the optimal Bayes solution (55.6) is that it requires knowledge of the probabilities π_r and $\mathbb{P}(\boldsymbol{h} = h|\boldsymbol{r} = r)$. The latter probabilities are determined once we know the parameters $\{p_{rm}\}$ under either the Bernoulli or multinomial model. The parameters $\{\pi_r, p_{rm}\}$ are rarely known beforehand and need to be estimated. We now assume that we have access to a collection of N training data points, $\{r(n), h_n,\, n = 0, 1, \ldots, N-1\}$. In this notation, $r(n)$ is the class for feature h_n.

Estimating the class priors

Assume that within the N data samples, there are N_r examples that belong to class r. Then, the derivation in Prob. 55.3 shows that π_r can, in principle, be estimated as follows:

$$\widehat{\pi}_r \;=\; \frac{N_r}{N} \tag{55.13}$$

which is the fraction of data points that belong to class r within the training set. However, an adjustment is needed to avoid situations where a particular class may not be represented in the training data, in which case we will end up with $\widehat{\pi}_r = 0$ for that r. To avoid this situation, it is customary to modify the above expression for estimating π_r by incorporating a form of smoothing known as *Laplace smoothing* — see Probs. 55.4 and 55.5. We extend N to $N + sR$, where we assume the presence of sR additional fictitious training samples. Here, the parameter s is positive and controls the amount of smoothing. The choice $s = 1$ is common and referred to as Laplace smoothing. Choices of $s < 1$ are referred to as Lidstone smoothing. Now, assuming the labels $r \in \{1, 2, \ldots, R\}$ are uniformly distributed within the sR virtual samples, then s of these samples will be expected to belong to each class. We then replace expression (55.13) for $\widehat{\pi}_r$ by

$$\boxed{\widehat{\pi}_r \;=\; \frac{N_r + s}{N + sR}, \quad r = 1, 2, \ldots, R} \qquad (\textbf{Laplace smoothing}) \tag{55.14}$$

Observe that when $s = 0$ we get $\widehat{\pi}_r = N_r/N$, and when $s \to \infty$ we get $\widehat{\pi}_r \to 1/R$. Therefore, the smoothing operation ensures that the estimate for π_r lies between the sample average (N_r/N) and the uniform probability $(1/R)$.

Estimating the reverse conditional probabilities

Similarly, we can use the training data to estimate the parameters $\{p_{rm}\}$. Consider first the multinomial case, where p_{rm} denotes the likelihood that attribute m occurs under class r. Given the N training feature vectors $\{h_n\}$, we isolate the vectors that belong to class r and count how many times attribute m occurs in them:

$$N_{rm} \;\overset{\Delta}{=}\; \sum_{h_n \in \text{ class } r} h_n(m) \tag{55.15a}$$

Note that m is fixed in this sum and we are adding over all feature vectors from class r in the training set. If we add the $\{N_{rm}\}$ over m, we arrive at the total number of all attributes observed in the training set under class r:

$$N_{rT} \;\overset{\Delta}{=}\; \sum_{m=1}^{M} N_{rm} \tag{55.15b}$$

Then, p_{rm} is estimated by using the smoothed formula

$$(\textbf{multinomial parameters})$$
$$\widehat{p}_{rm} \;=\; \frac{N_{rm} + s}{N_{rT} + sM}, \quad \left\{ \begin{array}{rl} m = & 1, \ldots, M \\ r = & 1, \ldots, R \end{array} \right. \tag{55.15c}$$

for some $s > 0$ since there are M possible attributes. This calculation assumes that the training data is dense enough so that all classes are observed.

For the Bernoulli model, we again isolate the vectors that belong to class r and count how many times attribute m is active at the value one within these vectors:

$$N_{rm} \;\overset{\Delta}{=}\; \sum_{h_n \in \text{ class } r} h_n(m) \tag{55.16a}$$

We also let N_r denote the total number of feature vectors in class r:

$$N_r = \text{number of features } h_n \text{ in class } r \tag{55.16b}$$

Then, p_{rm} is estimated by using the smoothed formula

$$\textbf{(Bernoulli parameters)}$$
$$\widehat{p}_{rm} \;=\; \frac{N_{rm}+s}{N_r+2s}, \qquad \left\{\begin{array}{rl} m = & 1,\ldots,M \\ r = & 1,\ldots,R \end{array}\right. \tag{55.16c}$$

The following listing summarizes the steps involved in the training and classification phases of the naïve Bayes classifier for multinomial-distributed feature data using (55.15c); for Bernoulli-distributed attributes, we use (55.16c) instead. The construction is relatively simple to train. Note that we are denoting the resulting classifier in the last line of the algorithm by the notation $r^\star(h)$ (as opposed to $r^\bullet(h)$) because it is learned directly from the training data.

Naive Bayes classifier for discrete multinomial feature data.

```
given $N$ training data points $\{r(n), h_n\}, n = 0, 1, 2, \ldots, N-1$;
given $R$ classes, $r(n) \in \{1, 2, \ldots, R\}$;
each feature vector, $h_n$, is $M-$dimensional with entries $\{h_n(m)\}$;
$h_n(m)$ counts how many times attribute $m$ occurs in $n-$th sample;
select a Laplace smoothing factor $s > 0$, e.g., $s = 1$.
(training)
repeat $r = 1, 2, \ldots, R$ :
    $N_r$ = number of training samples in class $r$;
    $\widehat{\pi}_r = \frac{N_r+s}{N+sR}$
    repeat $m = 1, 2, \ldots, M$ :
        $N_{rm} \overset{(55.15a)}{=}$ number of times attribute $m$ occurs in class $r$;
        $N_{rT} \overset{(55.15b)}{=}$ total number of attributes observed in class $r$;
        $\widehat{p}_{rm} = \dfrac{N_{rm}+s}{N_{rT}+sM}$
    end
end
(classification)
given a new feature vector, $h$, with entries $\{h(m)\}$ :
    compute $\widehat{\mathbb{P}}(\boldsymbol{h} = h|\boldsymbol{r} = r)$ using (55.12), for $r = 1, 2, \ldots, R$;
    determine $r^\star(h) = \underset{1\le r\le R}{\text{argmax}} \left\{\widehat{\pi}_r\, \widehat{\mathbb{P}}(\boldsymbol{h} = h|\boldsymbol{r} = r)\right\}$
end
```
(55.17)

Example 55.2 **(Application to medical diagnosis)** We reconsider the earlier Table 54.2 , repeated here, which lists the symptoms for $N = 10$ patients and whether they had the flu or not. e The number of classes in this example is $R = 2$ with:

$$\boldsymbol{\gamma} = +1 : \text{patient has the flu} \tag{55.18a}$$
$$\boldsymbol{\gamma} = -1 : \text{patient does not have the flu} \tag{55.18b}$$

The last column in the table indicates the class that each patient belongs to. Excluding this last column, each row in the table corresponds to a feature vector with $M = 6$ attributes. Each entry of $\boldsymbol{h}$ assumes a binary value (Yes/No); i.e., it is Bernoulli distributed. For example, the first entry of $\boldsymbol{h}$ indicates whether the patient had a headache or not. Figure 55.1 provides a graphical illustration of the data from Table 55.1, where the brown color indicates the presence of the relevant symptom. The top row in the figure lists patients without the flu, while the bottom row lists patients with the flu.

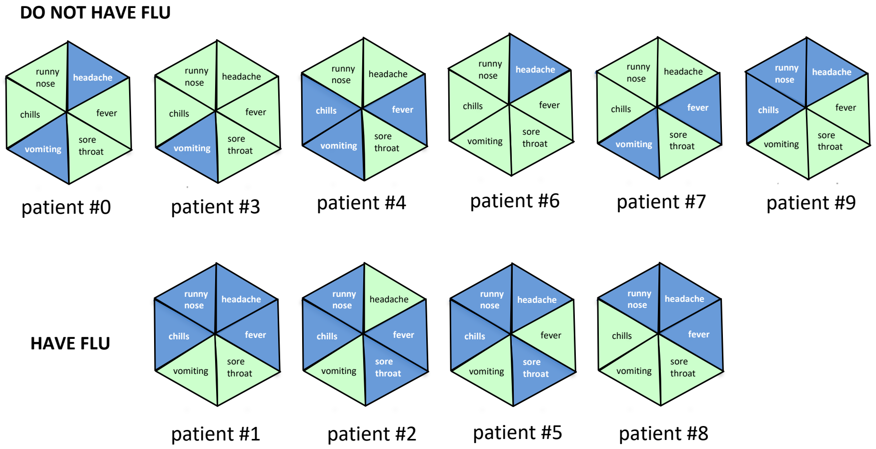

Figure 55.1 Graphical illustration of the data from Table 55.1, where the blue color indicates the presence of the relevant symptom. The top row lists patients without the flu, while the bottom row lists patients with the flu.

Table 55.1 Symptoms felt by 10 patients and whether they had the flu or not.

patient	headache	fever	sore throat	vomiting	chills	runny nose	FLU
0	Yes	No	No	Yes	No	No	NO
1	Yes	Yes	No	No	Yes	Yes	YES
2	No	Yes	Yes	No	Yes	Yes	YES
3	No	No	No	Yes	No	No	NO
4	No	Yes	No	Yes	Yes	No	NO
5	Yes	No	Yes	No	Yes	Yes	YES
6	Yes	No	No	No	No	No	NO
7	No	Yes	No	Yes	No	No	NO
8	Yes	Yes	No	No	No	Yes	YES
9	Yes	No	No	No	Yes	Yes	NO

We set the Laplace smoothing factor to $s = 1$ and use the data in the table to estimate the prior probabilities as follows:

$$N_{+1} = 4 \tag{55.19a}$$
$$N_{-1} = 6 \tag{55.19b}$$
$$\widehat{\pi}_{+1} = \frac{4+1}{10+2} \;\approx\; 0.4167 \tag{55.19c}$$
$$\widehat{\pi}_{-1} = \frac{6+1}{10+2} \;\approx\; 0.5833 \tag{55.19d}$$

where N_{+1} denotes the number of samples in the training set that belong to class $\gamma = +1$ (has the flu). Similarly for N_{-1}. We also use the data from the table to estimate the conditional probabilities, first for the patients that had the flu:

$$\widehat{\mathbb{P}}(\text{headache=yes}|\text{patient has flu}) = (3+1)/(4+2) = 2/3 \tag{55.20a}$$
$$\widehat{\mathbb{P}}(\text{headache=no}|\text{patient has flu}) = (1+1)/(4+2) = 1/3 \tag{55.20b}$$
$$\widehat{\mathbb{P}}(\text{fever=yes}|\text{patient has flu}) = (3+1)/(4+2) = 2/3 \tag{55.20c}$$
$$\widehat{\mathbb{P}}(\text{fever=no}|\text{patient has flu}) = (1+1)/(4+2) = 1/3 \tag{55.20d}$$
$$\widehat{\mathbb{P}}(\text{sore throat=yes}|\text{patient has flu}) = (2+1)/(4+2) = 1/2 \tag{55.20e}$$
$$\widehat{\mathbb{P}}(\text{sore throat=no}|\text{patient has flu}) = (2+1)/(4+2) = 1/2 \tag{55.20f}$$
$$\widehat{\mathbb{P}}(\text{vomiting=yes}|\text{patient has flu}) = (0+1)/(4+2) = 1/6 \tag{55.20g}$$
$$\widehat{\mathbb{P}}(\text{vomiting=no}|\text{patient has flu}) = (4+1)/(4+2) = 5/6 \tag{55.20h}$$
$$\widehat{\mathbb{P}}(\text{chills=yes}|\text{patient has flu}) = (3+1)/(4+2) = 2/3 \tag{55.20i}$$
$$\widehat{\mathbb{P}}(\text{chills=no}|\text{patient has flu}) = (1+1)/(4+2) = 1/3 \tag{55.20j}$$
$$\widehat{\mathbb{P}}(\text{runny nose=yes}|\text{patient has flu}) = (4+1)/(4+2) = 5/6 \tag{55.20k}$$
$$\widehat{\mathbb{P}}(\text{runny nose=no}|\text{patient has flu}) = (0+1)/(4+2) = 1/6 \tag{55.20l}$$

and similarly for the patients that did not have the flu:

$$\widehat{\mathbb{P}}(\text{headache=yes}|\text{patient does not have flu}) = (3+1)/(6+2) = 1/2 \tag{55.21a}$$
$$\widehat{\mathbb{P}}(\text{headache=no}|\text{patient does not have flu}) = (3+1)/(6+2) = 1/2 \tag{55.21b}$$
$$\widehat{\mathbb{P}}(\text{fever=yes}|\text{patient does not have flu}) = (2+1)/(6+2) = 3/8 \tag{55.21c}$$
$$\widehat{\mathbb{P}}(\text{fever=no}|\text{patient does not have flu}) = (4+1)/(6+2) = 5/8 \tag{55.21d}$$
$$\widehat{\mathbb{P}}(\text{sore throat=yes}|\text{patient does not have flu}) = (0+1)/(6+2) = 1/8 \tag{55.21e}$$
$$\widehat{\mathbb{P}}(\text{sore throat=no}|\text{patient does not have flu}) = (6+1)/(6+2) = 7/8 \tag{55.21f}$$
$$\widehat{\mathbb{P}}(\text{vomiting=yes}|\text{patient does not have flu}) = (4+1)/(6+2) = 6/8 \tag{55.21g}$$
$$\widehat{\mathbb{P}}(\text{vomiting=no}|\text{patient does not have flu}) = (2+1)/(6+2) = 3/8 \tag{55.21h}$$
$$\widehat{\mathbb{P}}(\text{chills=yes}|\text{patient does not have flu}) = (2+1)/(6+2) = 3/8 \tag{55.21i}$$
$$\widehat{\mathbb{P}}(\text{chills=no}|\text{patient does not have flu}) = (4+1)/(6+2) = 5/8 \tag{55.21j}$$
$$\widehat{\mathbb{P}}(\text{runny nose=yes}|\text{patient does not have flu}) = (1+1)/(6+2) = 1/4 \tag{55.21k}$$
$$\widehat{\mathbb{P}}(\text{runny nose=no}|\text{patient does not have flu}) = (5+1)/(6+2) = 3/4 \tag{55.21l}$$

In this example, we would like to employ the naïve Bayes classifier to decide whether a new patient with the following symptoms has the flu or not:

$$\begin{aligned} h = \{&\text{headache=NO, fever=NO, sore throat=YES,} \\ &\text{vomiting=NO, chills=NO, runny nose=YES}\} \end{aligned} \tag{55.22}$$

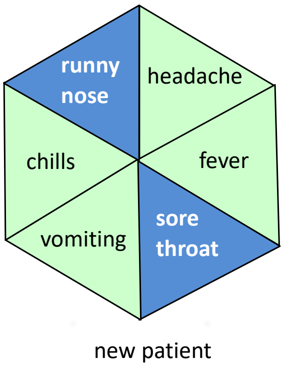

Figure 55.2 Graphical illustration of the symptoms for the new patient. Does the patient have the flu?

The symptoms for the new patient are represented graphically in Fig. 55.2. To begin with, using (55.5), we evaluate the following conditional probabilities:

$$\begin{aligned}
&\widehat{\mathbb{P}}(\boldsymbol{h} = h|\text{patient has flu})\\
&= \prod_{m=1}^{6} \widehat{\mathbb{P}}(\boldsymbol{h}(m) = h(m)|\text{patient has flu})\\
&= \widehat{\mathbb{P}}(\text{headache=no}|\text{patient has flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{fever=no}|\text{patient has flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{sore throat=yes}|\text{patient has flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{vomiting=no}|\text{patient has flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{chills=no}|\text{patient has flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{runny nose=yes}|\text{patient has flu})\\
&= 1/3 \times 1/3 \times 1/2 \times 5/6 \times 1/3 \times 5/6\\
&\approx 0.01286
\end{aligned} \tag{55.23}$$

and

$$\begin{aligned}
&\widehat{\mathbb{P}}(\boldsymbol{h} = h|\text{patient does not have flu})\\
&= \prod_{m=1}^{6} \widehat{\mathbb{P}}(\boldsymbol{h}(m) = h(m)|\text{patient does not have flu})\\
&= \widehat{\mathbb{P}}(\text{headache=no}|\text{patient does not have flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{fever=no}|\text{patient does not have flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{sore throat=yes}|\text{patient does not have flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{vomiting=no}|\text{patient does not have flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{chills=no}|\text{patient does not have flu}) \times\\
&\quad \widehat{\mathbb{P}}(\text{runny nose=yes}|\text{patient does not have flu})\\
&= 1/2 \times 5/8 \times 1/8 \times 3/8 \times 5/8 \times 1/4\\
&\approx 0.002289
\end{aligned} \tag{55.24}$$

Consequently,

$$\widehat{\pi}_{+1}\,\widehat{\mathbb{P}}(\boldsymbol{h} = h|\text{patient has flu}) \approx 0.4167 \times 0.01286 \quad \approx 0.005359 \tag{55.25}$$

and

$$\widehat{\pi}_{-1}\,\widehat{\mathbb{P}}(\boldsymbol{h} = h|\text{patient does not have flu}) \approx 0.5833 \times 0.002289 \approx 0.00013352 \tag{55.26}$$

Since $0.005359 > 0.00013352$, we conclude that the patient is likely to have the flu. This example helps illustrate one main limitation of naïve Bayes classifiers, namely, the assumption that the entries of the feature vector (i.e., the attributes) are conditionally independent of each other. For example, given that a patient has the flu, it is likely that having a fever and feeling a chill are dependent (rather than independent) events. Still, naïve Bayes classification is a popular learning scheme due to its computational simplicity and the fact that it performs surprisingly well (although it can be outperformed by other more elaborate learning methods).

Example 55.3 (**Application to spam filtering**) We apply the naïve Bayes classifier to another situation with $R = 2$ classes $\boldsymbol{\gamma} \in \{\pm 1\}$ (such as checking whether an email message is spam or not), with $\boldsymbol{\gamma} = +1$ corresponding to spam messages. In this example, each entry of the feature vector, $h \in \mathbb{R}^M$, is binary-valued and its value is either one or zero depending on whether a particular word is present in the message or not. Using Laplace smoothing, with $s = 1$, we first estimate the probabilities for the two classes:

$$\widehat{\pi}_{+1} = \frac{N_{+1}+1}{N+2}, \qquad \widehat{\pi}_{-1} = \frac{N_{-1}+1}{N+2} \tag{55.27}$$

where N_{+1} (similarly, N_{-1}) denotes the number of samples in the training set that belong to class $\boldsymbol{\gamma} = +1$ (similarly, $\boldsymbol{\gamma} = -1$). Likewise, for each $m = 1, 2, \ldots, M$, we used Laplace smoothing again to estimate the Bernoulli parameters:

$$\widehat{p}_{+1,m} \overset{\Delta}{=} = \widehat{\mathbb{P}}(\boldsymbol{h}(m) = 1|\boldsymbol{\gamma} = +1) = (N_{+1,m}+1)/(N_{+1}+2) \tag{55.28a}$$

$$\widehat{p}_{-1,m} \overset{\Delta}{=} \widehat{\mathbb{P}}(\boldsymbol{h}(m) = 1|\boldsymbol{\gamma} = -1) = (N_{-1,m}+1)/(N_{-1}+2) \tag{55.28b}$$

where

$$N_{+1,m} \overset{\Delta}{=} \text{number of observations in class } +1 \text{ having } h(m) = 1 \tag{55.29a}$$

$$N_{-1,m} \overset{\Delta}{=} \text{number of observations in class } -1 \text{ having } h(m) = 1 \tag{55.29b}$$

Using the above parameters we can write , for any $\gamma \in \{\pm 1\}$:

$$\widehat{\mathbb{P}}\Big(\boldsymbol{h}(m) = h(m)|\boldsymbol{\gamma} = \gamma\Big) = \widehat{p}_{\gamma,m}^{\,h(m)}\,(1-\widehat{p}_{\gamma,m})^{1-h(m)} \tag{55.30a}$$

Accordingly, given a new message with feature vector h, we can decide its class (whether spam or not) by seeking the value of $\gamma \in \{\pm 1\}$ that maximizes:

$$\gamma^\star(h) = \underset{\gamma\in\{\pm 1\}}{\text{argmax}} \Big\{\widehat{\pi}_\gamma\,\widehat{\mathbb{P}}(\boldsymbol{h} = h|\boldsymbol{\gamma} = \gamma)\Big\} \tag{55.31}$$

where

$$\widehat{\mathbb{P}}(\boldsymbol{h} = h|\boldsymbol{\gamma} = \gamma) \;=\; \prod_{m=1}^{M} \widehat{\mathbb{P}}\Big(\boldsymbol{h}(m) = h(m)|\boldsymbol{\gamma} = \gamma\Big) \tag{55.32}$$

55.4 GAUSSIAN NAÏVE CLASSIFIER

We have restricted so far the entries of the feature vector $\boldsymbol{h}$ to discrete values. The naïve Bayes construction can be extended to the case in which the entries of $\boldsymbol{h}$ are *continuous* in $\mathrm{I\!R}$. In this section, we describe the situation in which these entries continue to be conditionally independent of each other. Later, in Sec. 56.2, we will consider the more general scenario where the entries of $\boldsymbol{h}$ can be correlated and derive linear discriminant methods for approximating the Bayes classifier.

Let $\{\boldsymbol{h}(m)\}$ denote the individual entries of $\boldsymbol{h} \in \mathrm{I\!R}^M$. Assume that, conditioned on the class variable $\boldsymbol{r} = r$, each $\boldsymbol{h}(m)$ is Gaussian distributed with mean μ_{rm} and variance σ^2_{rm}, written as

$$\begin{aligned} f_{\boldsymbol{h}(m)|\boldsymbol{r}}(h(m)|r) &\sim \mathcal{N}_m(\mu_{rm}, \sigma^2_{rm}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2_{rm}}} \exp\left\{ -\frac{1}{2\sigma^2_{rm}} \Big(h(m) - \mu_{rm} \Big)^2 \right\} \end{aligned} \tag{55.33}$$

Note that we are using two subscripts to characterize the mean and variance parameters of the Gaussian distribution: the subscript r indicates that these parameters depend on the class label, and the subscript m refers to the $m-$th attribute. We are also denoting the Gaussian distribution for $\boldsymbol{h}(m)$ by the compact notation $\mathcal{N}_m$, with a subscript m. The independence assumption on the entries of $\boldsymbol{h}$ implies that

$$\begin{aligned} f_{\boldsymbol{h}|\boldsymbol{r}}(h|r) &= \prod_{m=1}^{M} \mathcal{N}_m(\mu_{rm}, \sigma^2_{rm}) \\ &= \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2_{rm}}} \exp\left\{ -\frac{1}{2\sigma^2_{rm}} (h(m) - \mu_{rm})^2 \right\} \end{aligned} \tag{55.34}$$

Repeating the argument that led to (55.4) using Bayes rule, we find that, given a feature vector h, the class selection $r^\bullet(h)$ can be determined by solving:

$$\begin{aligned} r^\bullet(h) &= \underset{1\le r\le R}{\text{argmax}} \;\; \pi_r \, f_{\boldsymbol{h}|\boldsymbol{r}}(h|r) \qquad\qquad (55.35) \\ &= \underset{1\le r\le R}{\text{argmax}} \;\; \left\{ \ln(\pi_r) - \frac{1}{2} \sum_{m=1}^{M} \Big(\ln(2\pi\sigma^2_{rm}) + \frac{1}{\sigma^2_{rm}} (h(m) - \mu_{rm})^2 \Big) \right\} \end{aligned}$$

The mean and variance parameters $\{\mu_{rm}, \sigma^2_{rm}\}$ can be estimated from the training data $\{r(n), h_n\}$. If we let N_r denote the number of feature vectors that belong to class r, then we set

$$\widehat{\mu}_{rm} = \frac{1}{N_r} \sum_{r(n)=r} h_n(m) \tag{55.36a}$$

$$\widehat{\sigma}^2_{rm} = \frac{1}{N_r - 1} \sum_{r(n)=r} \Big(h_n(m) - \widehat{\mu}_{rm}\Big)^2 \tag{55.36b}$$

The resulting algorithm is listed in (55.37).

Naive Bayes classifier for Gaussian feature data

```
given $N$ training data points $\{r(n), h_n\}, n = 0, 1, 2, \ldots, N-1$;
given $R$ classes, $r(n) \in \{1, 2, \ldots, R\}$;
each feature vector, $h_n$, is $M-$dimensional with entries $\{h_n(m)\}$;
$h_n(m)$ is Gaussian-distributed;
select a Laplace smoothing factor $s > 0$, e.g., $s = 1$.
(training)
repeat $r = 1, 2, \ldots, R$ :
    $N_r$ = number of training samples in class $r$;
    $\widehat{\pi}_r = \frac{N_r + s}{N + sR}$
    repeat $m = 1, 2, \ldots, M$ :
        $\widehat{\mu}_{rm} = \frac{1}{N_r} \sum_{r(n)=r} h_n(m)$
        $\widehat{\sigma}^2_{rm} = \frac{1}{N_r - 1} \sum_{r(n)=r} \Big(h_n(m) - \widehat{\mu}_{rm}\Big)^2$
    end
end
(classification)
given a new feature vector, $h$, with entries $\{h(m)\}$ :
    $r^\star(h) = \underset{1 \le r \le R}{\text{argmax}} \left\{ \ln(\widehat{\pi}_r) - \frac{1}{2} \sum_{m=1}^{M} \Big( \ln(2\pi\widehat{\sigma}^2_{rm}) + \frac{1}{\widehat{\sigma}^2_{rm}} (h(m) - \widehat{\mu}_{rm})^2 \Big) \right\}$
end
```

(55.37)

55.5 COMMENTARIES AND DISCUSSION

Laplace smoothing. The Laplace smoothing formula (55.14) is attributed to the French mathematician **Pierre-Simon Laplace (1749–1827)**. He derived it in the work by Laplace (1814) in his study of the *rule of succession*, which deals with the following question. Assume an experiment with only two possible outcomes (success or failure) is repeated a total of N independent times, and that N_s successes have been observed during these trials. Assume we only know that the experiment has two possible outcomes but have no information about the likelihood of each outcome. Consider now the question of determining the probability that the outcome will be a success in the $(N+1)-$th trial. This probability is given by — see, e.g., the textbooks by Doob (1953) and Jaynes (2003) and Probs. 55.4 and 55.5:

$$\mathbb{P}(\text{success in trial } N+1 \,|\, \text{given } N_s \text{ successes so far}) \;=\; \frac{N_s+1}{N+2} \tag{55.38}$$

This result can be extended to the case in which each trial has a total of R possible outcomes, say, $r \in \{1, 2, \ldots, R\}$. In this case, the probability that the outcome is in class r in the $(N+1)-$th trial will be given by:

$$\begin{aligned}&\mathbb{P}(\text{outcome is class } r \text{ in trial } N+1 \,|\, \text{given } N_r \text{ observations of } r \text{ so far})\\ &= \frac{N_r+1}{N+R}\end{aligned} \tag{55.39}$$

More generally, we can resort to expression (55.14) where $s > 0$. The choice $s = 1$ leads to Laplace smoothing, while choices $s < 1$ lead to Lidstone smoothing. Some of the earlier references on smoothing techniques include the works by Lidstone (1920), Johnson (1932), and Jeffreys (1948).

Naive Bayes classifier. According to Duda, Hart, and Stork (2000) and Russel and Norvig (2009), some of the earliest applications of the algorithm were in the context of pattern recognition, text classification, and medical diagnosis in the late 1950s and early 1960s. For example, the early work by Maron (1961) examines the task of automatically classifying documents into various categories; the author motivates the work in the abstract of the article by writing that *"the task, in essence, is to have a computing machine read a document and on the basis of the occurrence of selected clue words decide to which of many subject categories the document in question belongs."* The author motivates the naïve Bayes construction by using the Shannon entropy measure to quantify the uncertainty about which category a document belongs to.

We indicated in the body of the chapter that although the naïve Bayes classifier assumes the entries of the feature vector to be conditionally independent of each other, the classifier still performs competitively in practice even when the independence condition is violated. There have been several studies in the literature to illustrate and explain this behavior, most notably by Clark and Niblett (1989), Langley, Iba, and Thompson (1992), Kononenko (1993), Pazzani (1996), Domingos and Pazzani (1996, 1997), Frank *et al.* (2000), Garg and Roth (2001), Hand and Yu (2001), and Zhang (2004). The main conclusion from these works is that while the estimates of the conditional probabilities, $\widehat{\mathbb{P}}(\boldsymbol{r} = r|\boldsymbol{h} = h)$, can generally be poor (i.e., not close enough to their true values), the naïve classifier is still able to deliver performance because the predicted class, $r^\star$, is decided not based on the estimated values of the probabilities but rather on comparing these values against each other, i.e., on selecting the class $r^\star$ that leads to the largest value for $\widehat{\mathbb{P}}(\boldsymbol{r} = r|\boldsymbol{h} = h)$.

Naïve Bayes classifiers can be outperformed by other learners as shown, for example, in the works by Ng and Jordan (2001) and Caruana and Niculescu-Mizil (2006). The first work compared logistic regression and naïve Bayes, while the second work compared several learning algorithms against each other including logistic regression, support vector machines, and naïve Bayes. Nevertheless, motivated by the extensive empirical and analytical evidence in support of the good performance of naïve Bayes classifiers in many situations of interest, these classifiers continue to serve as good starting points for the design of more elaborate learning machines.

PROBLEMS

55.1 Repeat the derivation of Example 55.2 to verify whether a patient with the following symptoms has the flu:

$$\begin{aligned} h = \{&\text{headache=YES, fever=YES, sore throat=NO,} \\ &\text{vomiting=NO, chills=NO, runny nose=NO}\} \end{aligned}$$

55.2 Continuing with the same patient from the previous problem, assume the feature vector is missing information about whether the patient has a sore throat or not (marked by the question mark below):

$$\begin{aligned} h = \{&\text{headache=YES, fever=YES, sore throat=?,} \\ &\text{vomiting=NO, chills=NO, runny nose=NO}\} \end{aligned}$$

How would you apply the naïve Bayes classifier to decide on whether the patient has the flu or not? Assuming the patient had a 60% chance of having a sore throat, how likely is it that the decision based on ignoring this information will be different from the decision that takes this additional piece of information into consideration?

55.3 Consider a multiclass classification problem consisting of R classes, say, $\boldsymbol{r} \in \{1, 2, \ldots, R\}$. The prior probability of observing features from class $\boldsymbol{r}$ is denoted by π_r. A collection of N independent realizations $\{\boldsymbol{r}(n), \boldsymbol{h}_n\}$ are observed, with $\boldsymbol{r}(n)$ denoting the class variable and $\boldsymbol{h}_n$ the corresponding feature vector for the $n-$th sample. It is observed that each class r occurs N_r times in the sample of N data points.

(a) Determine the likelihood probability $\mathbb{P}(N_1, N_2, \ldots, N_R | \pi_1, \pi_2, \ldots, \pi_R)$, where the $\{\pi_r\}$ are treated as deterministic parameters.

(b) Show that the optimal estimate for π_r that is obtained by maximizing the logarithm of the above probability expression is given by $\widehat{\pi}_r = N_r/N$.

55.4 One way to motivate expression (55.14) for Laplace smoothing is as follows. We continue with the setting of Prob. 55.3 except that we now model the unknown priors $\{\boldsymbol{\pi}_r\}$ as random variables whose individual pdfs follow a symmetric Dirichlet distribution with parameter $s > 0$. Since the $\{\boldsymbol{\pi}_r\}$ should add up to one, this means that one of the variables is fully determined from knowledge of the remaining $R - 1$ variables. A joint Dirichlet pdf with positive parameters $\{s_1, s_2, \ldots, s_R\}$ has the form:

$$f_{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_R}(\pi_1, \pi_2, \ldots, \pi_R) \;\propto\; \prod_{r=1}^{R} \pi_r^{s_r - 1}$$

where the symbol $\propto$ denotes proportionality. It is known that the mean of each entry $\boldsymbol{\pi}_r$ under this distribution is given by $\mathbb{E}\,\boldsymbol{\pi}_r = s_r / \sum_{r=1}^{R} s_r$. When $s_r = s$, for all $r \in \{1, 2, \ldots, R\}$, the distribution is said to be symmetric.

(a) Verify that $\mathbb{P}(N_1, \ldots, N_R | \boldsymbol{\pi} = \pi) \propto \prod_{r=1}^{R} \pi_r^{N_r}$.

(b) Assuming a symmetric distribution, verify that

$$f_{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_R}(\pi_1, \pi_2, \ldots, \pi_R | N_1, \ldots, N_R) \;\propto\; \prod_{r=1}^{R} \pi_r^{N_r + s - 1}$$

(c) Conclude that the optimal mean-square-error estimate for $\boldsymbol{\pi}_r$ given the observations $\{N_1, N_2, \ldots, N_R\}$, which is equal to the expectation of the conditional pdf of part (b), is given by:

$$\widehat{\pi}_r \;=\; \frac{N_r + s}{N + sR}$$

55.5 Derive Laplace formula (55.38). Using this formula, what would the probability of the sun rising tomorrow be? Any controversy in the answer? *Remark.* This sun problem was used by Laplace (1814) to illustrate his calculation.

55.6 Refer to expression (55.12) when the entries of $\boldsymbol{h}$ follow a multinomial distribution. Show that the Bayes classifier (55.6) reduces to the following equivalent problem involving an affine function of the feature data:

$$r^{\bullet}(h) = \underset{r\in\{1,2,\ldots,R\}}{\text{argmax}} \left\{\log(\pi_r) + h^{\mathsf{T}} w_r\right\}$$

where $w_r \in \mathbb{R}^M$ collects the log values of the attribute probabilities:

$$w_r \triangleq \begin{bmatrix} \log(p_{r1}) & \log(p_{r2}) & \ldots & \log(p_{rM}) \end{bmatrix}$$

55.7 Refer again to expression (55.12) when the entries of $\boldsymbol{h}$ follow a multinomial distribution. Assume there are two classes, $R = 2$, denoted by $\gamma \in \{\pm 1\}$. Show that the Bayes classifier (55.6) reduces to checking the sign of an affine function of the feature data as follows:

$$r^{\bullet}(h) = \text{sign}(h^{\mathsf{T}} w^{\bullet} - \theta^{\bullet})$$

where the parameters are given by

$$\theta^{\bullet} = \ln(\pi_{-1}/\pi_{+}), \quad w^{\bullet} = \text{col}\Big\{\ln(p_{+1,m}/p_{-1,m})\Big\} \in \mathbb{R}^M$$

55.8 For the naïve Bayes classifier, how many conditional probabilities of the form (55.15c) need to be estimated from the training data?

55.9 Refer to expression (52.8) for the Bayes classifier. Assume the $j-$th entry of the feature vector h is missing at random, denoted by h_j. Let h_{-j} denote the remaining entries of h; it is a vector of size $M-1$. Let $r^{\bullet}(h_{-j})$ denote the optimal class label based on knowledge of h_{-j} alone. Under the independence assumption, argue that

$$r^{\bullet}(h_{-j}) \;=\; \underset{1\leq r\leq R}{\text{argmax}} \left\{\pi_r \times \prod_{i\neq j} \mathbb{P}(\boldsymbol{h}_i = h_i | \boldsymbol{r} = r)\right\}$$

so that classification can proceed by ignoring h_j.

55.10 Refer to expressions (55.36a)–(55.36b) for estimating the parameters of a Gaussian naïve classifier. How may parameters $\{\mu_{rm}, \sigma_{rm}^2\}$ need to be estimated in total?

55.11 Assume the variances $\{\sigma_{rm}^2\}$ in the Gaussian naïve implementation are independent of the class label r and can be replaced by the notation $\{\sigma_m^2\}$. How would you estimate the $\{\sigma_m^2\}$?

55.12 Refer to expression (55.34) for the conditional pdf of $\boldsymbol{h}$ given the class variable in the Gaussian naïve classifier. Assume there are two classes denoted by $\gamma \in \{\pm 1\}$ with priors $\{\pi_{+1}, \pi_{-1}\}$. Assume further that the variances σ_{rm}^2 are independent of r and denote them by σ_m^2. Show that the conditional probability $\mathbb{P}(\boldsymbol{\gamma} = \gamma | \boldsymbol{h} = h)$ can be written in the following sigmoidal form

$$\mathbb{P}(\boldsymbol{\gamma} = \gamma | \boldsymbol{h} = h) = \frac{1}{1 + e^{-\gamma(h^{\mathsf{T}} w - \theta)}}$$

for some parameters (w, θ). Determine expressions for these parameters in terms of $\{\mu_{+1,m}, \mu_{-1,m}, \sigma_m^2, \pi_{+1}, \pi_{-1}\}$.

REFERENCES

Clark, P. and T. Niblett (1989), “The CN2 induction algorithm,”*Machine Learning*, vol. 3, no. 4, pp. 261–283.

Caruana R. and A. Niculescu-Mizil (2006), “An empirical comparison of supervised learning algorithms,” *Proc. Inter. Conf. on Machine Learning* (ICML), pp. 161–168, Pittsburgh, PA.

Domingos, P. and M. Pazzani (1996), “Beyond independence: Conditions for the optimality of the simple Bayesian classifier,” *Proc. International Conference on Machine Learning* (ICML), pp. 1–8, Bari, Italy.

Domingos, P. and M. Pazzani (1997), “On the optimality of the simple Bayesian classifier under zero-one loss,” *Machine Learning*, vol. 29, pp. 103–130.

Doob, J. L. (1953), *Stochastic Processes*, Wiley, NY.

Duda, R. O., P. E. Hart, and D. G. Stork (2000), *Pattern Classification*, 2nd edition, Wiley, NY.

Frank, E., L. Trigg, G. Holmes, and I. H. Witten (2000), “Naïve Bayes for regression,” *Machine Learning*, vol. 41, no. 1, pp. 5–15.

Garg, A. and D. Roth (2001), “Understanding probabilistic classifiers,” In L. D. Raedt and P. Flach, *Eds.*, *Proc. European Conference on Machine Learning,* Springer, pp. 179–191.

Hand, D. J. and Y. Yu (2001), “Idiot’s Bayes – not so stupid after all?” *International Statistical Review*, vol. 69, pp. 385–389.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press.

Jeffreys, H. (1948), *Theory of Probability*, 2nd edition, Clarendon Press, Oxford.

Johnson, W. E. (1932), “Probability: Deductive and inductive problems,” *Mind*, vol. 41, pp. 421–423.

Kononenko, I. (1993), “Inductive and Bayesian learning in medical diagnosis,” *Applied Artificial Intelligence*, vol. 7, pp. 317–337.

Langley, P., W. Iba, and K. Thompson (1992), “An analysis of Bayesian classifiers,” *Proc. National Conference on Artificial Intelligence* (AAAI), pp. 223–228, San Jose, CA.

Laplace, P. S. (1814), *Essai Philosophique sur les Probabilités*, Paris, published 1840.

Lidstone, G. J.(1920), “Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities,” *Transactions of the Faculty of Actuaries*, vol. 8, pp. 182–192.

Maron, M. E. (1961), “Automatic indexing: An experimental inquiry,” *Journal of the ACM*, vol. 8, no. 3, pp. 404–417.

Ng, A. Y. and M. I. Jordan (2001), “On discriminative vs. generative classifiers: A comparison of logistic regression and naïve Bayes,” *Proc. Advances Neural Information Systems* (NIPS), pp. 1–8, Vancouver, Canada.

Pazzani, M. (1996), “Searching for dependencies in Bayesian classifiers,” in D. Fisher D. and H. J. Lenz, *Eds.*, *Learning from Data.* Lecture Notes in Statistics, vol 112. Springer, NY.

Russell, S. and P. Norvig (2009), *Artificial Intelligence: A Modern Approach*, 3rd edition, Prentice Hall, NJ.

Zhang, R. (2004), “The optimality of naïve Bayes,” *Proc. AAAI Int. FLAIRS Conference*, pp. 1–6, Miami Beach, FL.