CHAPTER 4

The Bayes Classifier

Learning Objectives

At the end of this chapter, you will be able to:

- Explain the Bayes classifier
- Describe probability, conditional probability and Bayes' rule
- Define random variables, probability mass function, probability density function, cumulative distribution function, expectation and variance
- Explain optimality of the Bayes classifier
- Describe parametric and non-parametric schemes for density estimation
- Define class conditional independance and the naïve Bayes classifier

4.1 Introduction to the Bayes Classifier

The Bayes classifier is an optimal classifier. It operates based on the probability structure associated with the domain of application. It employs Bayes' rule to convert the *a priori* probability of a class into posterior probability with the help of probability distributions associated with the class. These posterior probabilities are used to classify the test patterns; the pattern is assigned to the class that has the largest posterior probability. In case the probability structure is not readily available, the training data is used to learn the probability structure.
Some of the important properties of Bayesian classifiers are as follows:

- It minimizes the probability of error associated with classification.
- It can deal with data that employs both categorical and numerical features.
- It is more of a benchmark classifier having sound theoretical properties. However, in most practical applications, the underlying probability structure is not readily available.
- Some additional constraints on the probability structure are applied to make estimation of the probability structure simpler. For example, the **naïve Bayes classifier (NBC)** is one that assumes that features are independent of each other, given that the data points belong to a class.
- Primarily, there are two schemes for estimating the probabilities associated. One of them depends solely on the data; it is called the maximum likelihood estimate or the frequency-based estimate. The other scheme is more general and it combines application domain knowledge with the data in estimation; it is called the **Bayesian estimation or Bayesian learning** of the probability structure.

4.2 Probability, Conditional Probability and Bayes' Rule

Let us start with a simple scenario where domain knowledge is used in the form of prior probabilities to classify patterns.

Example 1: Let us assume that 10,000 people in a community had undergone the COVID-19 test and 50 of them tested positive, while the remaining 9950 tested negative. In this two-class (binary class) problem, a simple frequency estimate will give us the following values for the probabilities for the two classes:

$$P(positive) = \frac{50}{10000} = 0.005 \text{ and } P(negative) = \frac{9950}{10000} = 0.995$$

These probabilities are called **prior probabilities** as they are obtained using domain knowledge.

Suppose a new person, from the community, who has not undergone the test, needs to be classified. In the absence of any other information from the person, one would try to use the prior probabilities in decision making; so the new person is assigned a COVID-19-negative label as the probability *P(negative)* is significantly larger than *P(positive)* ($0.995 >> 0.005$). So, invariably every other person in the community who has not undergone the test will be classified as COVID-19-negative using this rule of classification, which is influenced by the larger prior probability of being COVID-19-negative.

Such a classification is erroneous if the new person is actually COVID-19-positive. This can occur with a probability of 0.005 (*P(positive)*); so, the *probability of error* is 0.005. Note that every person who is actually COVID-19-negative is correctly classified in this process.

This is not new; in the KNN classifier, if the value of $k = n$, where n is the size of the training data set, and if we have k_1 (out of k) from the COVID-19-positive class and k_2 ($k - k_1$) from the COVID-19-negative class, then the probability estimates are

$$P(positive) = \frac{k_1}{n} \text{ and } P(negative) = \frac{k_2}{n}$$

It is not difficult to see that KNN gives the same result as classification based on prior probabilities.

One can ask whether we can do better if more information is available. In order to answer this question, we need to refresh some basic probability concepts:

- In a random experiment, we have a **sample space**, S, that is, the set of all outcomes. For example, tossing a coin gives us {H,T} as the sample space, where H stands for *head* and T stands for *tail*.
- An **event** is a subset of the sample space. We associate probabilities with events. If A is an event, then its probability $P(A) \in [0, 1]$, that is, probability is non-negative and is upper bounded (less than or equal to) 1.
- If A and B are disjoint events ($A \cap B = \phi$), where $A \cap B$ is the intersection of the sets A and B and ϕ is the null set or empty set, then

$$P(A \cup B) = P(A) + P(B),$$

where $A \cup B$ is the union of the sets A and B. This property holds for a countable union of events if they are pairwise disjoint.

EXAMPLE 2: If a coin is tossed twice, the sample space is {HH,HT,TH,TT}. If $A = \{HH,HT\}$ and $B = \{TT\}$, then $A \cap B = \phi$. Further, $P(A) = \frac{2}{4}$ as A has 2 out of the 4 outcomes, and $P(B) = \frac{1}{4}$. Note that $A \cup B = \{HH, HT, TT\}$; so,

$$P(A \cup B) = \frac{3}{4} = P(A) + P(B) = \frac{2}{4} + \frac{1}{4}$$

However, if $A = \{HH,HT\}$ and $C = \{HT, TT\}$, then $A \cap C = \{HT\} \neq \phi$. So, $P(A \cup C) = \frac{3}{4}$ whereas $P(A) = \frac{2}{4}$ and $P(C) = \frac{2}{4}$. So, here $P(A \cup C) \neq P(A) + P(C)$. It is possible to show that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If an event is described as *at least one tail*, then the event is {HT,TH,TT} and the probability of the event is $\frac{3}{4}$, as out of 4 elements of the sample space, 3 elements are favourable to the event.

Given an event A, its compliment A^c is given by the set difference $S - A$. Figure 4.1 shows regions corresponding to various related events, given two events A and B. Note that $A \cap B^c$ is the intersection of events A and B^c. The region for $A \cap B$ is the intersection of events A and B. The event $A^c \cap B$ corresponds to the intersection of the events A^c and B. Finally the remaining region indicates the intersection of the events A^c and B^c.

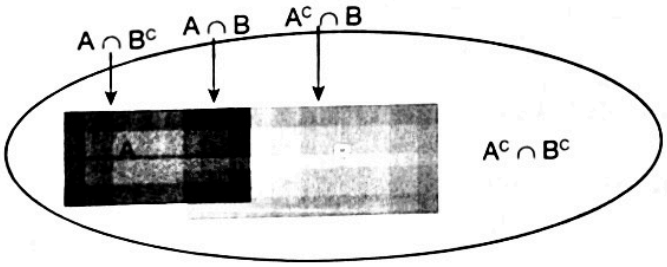

FIG. 4.1 Some events related to two given events A and B (for colour figure, please see Colour Plate 1)

4.2.1 Conditional Probability

We need to update the probability values if new information is given. Consider the following example.

EXAMPLE 3: Consider tossing a coin twice. We have seen in Example 2 that the probability of at least one tail is $\frac{3}{4}$. The corresponding event is {HT, TH, TT}. If we are given additional information that one of the tosses has resulted in a head, then the sample space is constrained to {HT, HH, TH}. So, under the condition that one toss has resulted in a head, the sample space shrinks. Now for at least one tail, the event in the new sample space is {HT,TH}. So, the probability has reduced from $\frac{3}{4}$ to $\frac{2}{3}$ for the event in the presence of more information. This computation is captured by using the notion of **conditional probability**.

If A and B are two events such that $P(B) \neq 0$, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

It is the probability of A conditioned on B. It is undefined if $P(B) = 0$. In the current example if A is the event specified by *at least one tail* and B is the event specified by *one toss that has resulted in a head*, then $B = \{\text{HH, HT, TH}\}$; so, $P(B) = \frac{3}{4}$. Note that $A \cap B = \{\text{TH,HT}\}$. So,

$$P(A|B) = \frac{\frac{2}{4}}{\frac{3}{4}} = \frac{2}{3}$$

as computed earlier in this example.

We have another important concept called **independent events**. We say that two events A and B are independent if $P(A \cap B) = P(A) \times P(B)$. So, if A and B are independent, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

4.2.2 Total Probability

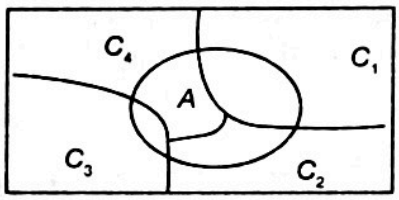

FIG. 4.2 An example to illustrate total probability

Consider the Venn diagram shown in Fig. 4.2. Here, event A is represented by the elliptical region and there are 4 events C_1, C_2, C_3 and C_4 that overlap with A. We can represent A using these overlapping sets by

$$A = (A \cap C_1) \cup (A \cap C_2) \cup (A \cap C_3) \cup (A \cap C_4)$$

The four overlapping sets in the union are disjoint as C_1, C_2, C_3 and C_4 are disjoint. Let

$$B_i = A \cap C_i \text{ for } i = 1, 2, 3, 4$$

So, $P(B_i) = P(A|C_i)P(C_i)$. So,

$$P(A) = P(B_1) + P(B_2) + P(B_3) + P(B_4)$$

Hence,

$$P(A) = P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + P(A|C_3)P(C_3) + P(A|C_4)P(C_4)$$

We know that $P(C_i|A) = \frac{P(A|C_i)P(C_i)}{P(A)}$. So, in general

$$P(C_i|A) = \frac{P(A|C_i)P(C_i)}{P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + P(A|C_3)P(C_3) + P(A|C_4)P(C_4)}$$

4.2.3 Bayes' Rule and Inference

Let us examine Bayes' rule and inference based on conditional probabilities. We consider some of the terms:

- $P(C_i)$: Prior or initial probability of event C_i.
- Model of the world (A) under each C_i: $P(A|C_i)$
- How to infer $P(C_i|A)$, or what is the probability $P(C_i|A)$ or equivalently how the *prior probability* $P(C_i)$ gets updated to the *posterior probability* $P(C_i|A)$. Let us consider an example to appreciate these ideas further.

EXAMPLE 4: Let C_1 and C_2 be two chests such that C_1 has 20 white balls (WB) and 10 red balls (RB); C_2 has 15 WBs and 15 RBs.
If one of the two chests is picked with equal probability and a ball randomly picked from the chest is WB, what is the probability that it came from C_1? We have the following information:

- Prior: $P(C_1) = P(C_2) = \frac{1}{2}$
- Given: $P(WB|C_1) = \frac{2}{3}$, $P(RB|C_1) = \frac{1}{3}$ and $P(WB|C_2) = P(RB|C_2) = \frac{1}{2}$
- Needed: $P(C_1|WB) = \frac{P(WB|C_1)P(C_1)}{P(WB|C_1)P(C_1)+P(WB|C_2)P(C_2)}$
  $= \frac{(\frac{2}{3})(\frac{1}{2})}{(\frac{2}{3})(\frac{1}{2})+(\frac{1}{2})(\frac{1}{2})} = \frac{4}{7}$

Let us consider one more example.

EXAMPLE 5: A new COVID-19 test claims to have 90% **true positive rate** (sensitivity) and 98% **true negative rate** (specificity). In a population with a COVID-19 prevalence of $\frac{1}{1000}$ (one out of 1000), what is the chance that a patient who tested positive is truly positive?
Let us consider the following:

- Let A be the event that a patient is **truly positive**; So, $P(A) = 0.001$.
- So, A^c $(S - A)$ is the event of being **truly negative**. So, $P(A^c) = 0.999$.
- Let B be the event that the patient **tested positive**.
- We want $P(A|B)$. Let the data of sensitivity and specificity be summarized as shown in Table 4.1.

TABLE 4.1 Relevant probabilities of positive and negative tests

True/Test	Positive	Negative
Positive	0.9	0.1
Negative	0.02	0.98

- The rows in Table 4.1 correspond to being truly positive and truly negative whereas the columns signify tested positive and tested negative.
- Note that the probability of being tested positive when truly positive is 0.9 (true positive rate); so, the probability of being tested negative when truly positive is 0.1.
- Similarly, in the second row, the entry 0.98 is the probability of being tested negative when truly negative (true negative rate) and 0.02 is the probability of being tested positive when the patient is truly negative.

We need to compute $P(A|B)$; we can use Bayes' rule to obtain $P(A|B)$ given $P(B|A)$, $P(A)$, $P(B|A^c)$ and $P(A^c)$ as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

$$= \frac{0.9 \times 0.001}{0.9 \times 0.001 + 0.02 \times 0.999} = 0.0431$$

So, more than 95% of those testing positive will be actually negative.

We can compute conditional probabilities using the chain rule. The chain rule is typically based on multiple applications of Bayes' rule.

Consider $P(A|B,C)$ for events A, B and C. It may be computed using the following chain rule:

$$P(A|B,C) = \frac{P(A) \times P(B,C|A)}{P(B,C)} = \frac{P(A) \times P(B|A) \times P(C|A,B)}{P(B,C)}$$

Similarly,

$$P(A^c|B,C) = \frac{P(A^c) \times P(B,C|A^c)}{P(B,C)} = \frac{P(A^c) \times P(B|A^c) \times P(C|A^c,B)}{P(B,C)}$$

So, if we compare $P(A|B,C)$ and $P(A^c|B,C)$, we need to consider only the numerators as the denominators of both the quantities are equal to $P(B,C)$. In such cases, we can write

$$P(A|B,C) \propto P(A) \times P(B|A) \times P(C|A,B)$$

and

$$P(A^c|B,C) \propto P(A^c) \times P(B|A^c) \times P(C|A^c,B)$$

4.2.4 Bayes' Rule and Classification

Let C_1 and C_2 be two classes, with their respective prior probabilities being $P(C_1)$ and $P(C_2)$. Given an object x, we can compute the posterior probabilities using Bayes' rule as follows:

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

and

$$P(C_2|x) = \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

We assign x to class C_1 if $P(C_1|x) > P(C_2|x)$, else we assign x to class C_2. What we are essentially doing is to assign the test pattern x to the class with the larger posterior probability. We illustrate this with the help of a simple example.

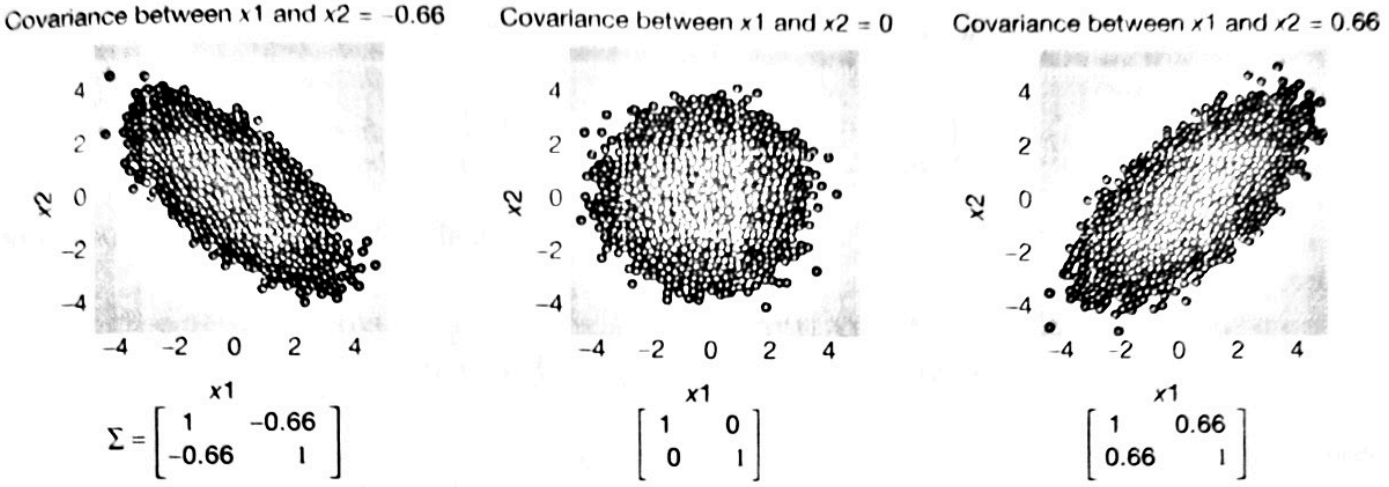

FIG. 4.10 Normally distributed two-dimensional points (for colour figure, please see Colour Plate 1)

Note that when Σ is a diagonal matrix, we have a circular shape for the distribution of points (Fig. 4.10 (b)). When the off-diagonal entries are non-zero, we have elliptical regions (Fig. 4.10 (a) and (c)) with different orientations based the polarity of these entries.

4.5 THE BAYES CLASSIFIER AND ITS OPTIMALITY

We have seen in Section 4.2.4 how posterior probabilities are obtained from prior probabilities. We have seen that

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x)} \Rightarrow \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Note that $P(x) = P(x|C_1)P(C_1) + P(x|C_2)P(C_2)$; it is a normalizer to ensure that $P(C_1|x) + P(C_2|x) = 1$.

The Bayes classifier assigns a test pattern x to C_1 if $P(C_1|x) > P(C_2|x)$, else to C_2. So, for a given x, the **probability of error** is

$$P(error|x) = min(P(C_1|x), P(C_2|x))$$

So, the **average or expected error** across all possible values of x is

$$\int_x P(error|x)f(x)dx = \int_x min(P(C_1|x), P(C_2|x))f(x)dx$$

Here, $f(x)$ is the PDF of x and it is fixed; note that for every x, we take a decision so that $P(error|x)$ is minimum. So, the Bayes classifier is optimal in the sense that it minimizes the average probability of error or error rate. So, it is the *minimum error rate classifier*.

Note that

$$P(C_1|x) > P(C_2|x) \Rightarrow P(x|C_1)P(C_1) > P(x|C_2)P(C_2)$$

We consider some examples to illustrate the use of the Bayes classifier.

EXAMPLE 15: Consider two classes defined in terms of how the values of x are distributed in each class as follows:

$$P(x|C_1) = \begin{cases} \frac{1}{2} & 0 \leq x \leq 2 \\ 0 & else \end{cases} \quad P(x|C_2) = \begin{cases} \frac{1}{5} & 1 \leq x \leq 6 \\ 0 & else \end{cases}$$

Let us assume that $P(C_1) = P(C_2) = 0.5$; the priors are equal. Let $x = 2$. Then the posterior values

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)} = \frac{\frac{1}{2} \times (0.5)}{\frac{1}{2} \times (0.5) + \frac{1}{5} \times (0.5)} \approx 0.7$$

and

$$P(C_2|x) = \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)} = \frac{\frac{1}{5} \times (0.5)}{\frac{1}{2} \times (0.5) + \frac{1}{5} \times (0.5)} \approx 0.3$$

So, the pattern x with a value of 2 is assigned to class C_1 by comparing the posteriors as $P(C_1|x)$ $P(C_2|x)$.

EXAMPLE 16: Let two classes be normally distributed with the same variance $\sigma^2 = 1$. Let the mean μ_1 of Class 1 be 2 and the mean of Class 2, μ_2, be 4. Let the prior probabilities be equal for the two classes, that is, $P(C_1) = P(C_2) = 0.5$. Let the test pattern be $x = 2$. Note that we can compare either the posterior probabilities or the numerators of the posteriors as the denominators are the same for both the posteriors. So, we consider the numerators in the posteriors for C_1 and C_2 (they are $f(x|C_1)P(C_1)$ and $f(x|C_2)P(C_2)$. These numerator quantities are as follows:

$f(x = 2|C_1) \times P(C_1) = \frac{1}{\sqrt{2\pi 1^2}} e^{-\frac{1}{2}\frac{(2-2)^2}{1^2}} \times 0.5 = \frac{0.5}{\sqrt{2\pi}}$

$f(x = 2|C_2) \times P(C_2) = \frac{1}{\sqrt{2\pi 1^2}} e^{-\frac{1}{2}\frac{(2-4)^2}{1^2}} \times 0.5 = \frac{0.5}{\sqrt{2\pi e^2}}$

So, $P(C_1|x) \propto \frac{0.5}{\sqrt{2\pi}}$ and $P(C_2|x) \propto \frac{0.5}{\sqrt{2\pi e^2}} \Rightarrow P(C_1|x) > P(C_2|x)$

So, assign the test pattern $x = 2$ to Class C_1.

It is possible to extend these ideas to d-dimensional vectors ($d > 1$).
Recall that multi-variate normal, for example, is characterized by **mean vector** μ and the **covariance matrix** Σ. If the vectors are d-dimensional, then μ is a d-dimensional vector and Σ is a $d \times d$ symmetric matrix, that is, $\Sigma_{i,j} = \Sigma_{j,i}$. All the diagonal entries are variances: $\Sigma_{i,i}$ is the variance of the i^{th} feature; $\Sigma_{i,j}$ is the covariance between the i^{th} and j^{th} features.
So, the decision making in the d-dimensional case is as follows:

- If $f(x|C_1)P(C_1) > f(x|C_2)P(C_2)$, assign x to C_1, else assign x to C_2.
- If $P(C_1) = P(C_2)$, we need to compare only the likelihood values $f(x|C_1)$ and $f(x|C_2)$.
- If the covariance matrices are equal, that is, $\Sigma_1 = \Sigma_2 = \sigma^2 I$, then the covariance matrices are diagonal and all the diagonal entries are equal to σ^2.
- Under the given conditions, $f(x|C_1) > f(x|C_2) \Rightarrow e^{-\frac{1}{2}(x-\mu_1)^t \frac{I}{\sigma^2}(x-\mu_1)} > e^{-\frac{1}{2}(x-\mu_2)^t \frac{I}{\sigma^2}(x-\mu_2)} \Rightarrow (x-\mu_1)^t(x-\mu_1) > (x-\mu_2)^t(x-\mu_2)$.
  This means assign x to C_1 if the squared Euclidean distance between x and μ_1 is less than the squared Euclidean distance between x and μ_2; equivalently assign x to that class whose mean is closer to x based on Euclidean distance.

This is depicted in Fig. 4.11. Note that the decision boundary (the broken line) that separates the two classes is the perpendicular bisector of the line joining the two means. Any test pattern x falling to the left of the decision boundary is classified as belonging to C_1; points on the right-hand side of the decision boundary are classified as belonging to Class C_2.

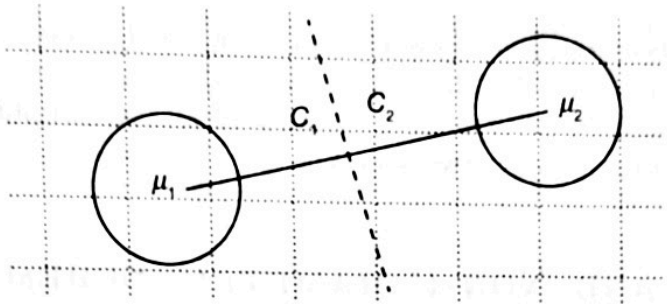

FIG. 4.11 Minimal distance classifier

This simple classifier is called the **minimal distance classifier (MDC)** and it is optimal when the priors are equal and the classes are normally distributed with the same covariance matrix; the covariance matrix is diagonal with the same entries in the diagonal locations.
If $\Sigma_1 = \Sigma_2$, it is possible to show that assigning x to C_1 is optimal if the squared Mahalanobis distance between x and μ_1 is smaller than that between x and μ_2.

4.5.1 Multi-Class Classification

We have discussed the use of the Bayes classifier in the two-class case. It can be easily used to deal with multi-class cases, that is, when the number of classes is more than 2. It may be described as follows:

- Let the classes be $C_1, C_2, \cdots, C_q$, where $q \geq 2$.
- Let the prior probabilities be $P(C_1), P(C_2), \cdots, P(C_q)$.
- Let x be the test pattern to be classified as belonging to one of these q classes.
- Compute the posterior probabilities using Bayes' rule

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{\sum_{j=1}^{q} P(x|C_j)P(C_j)}, \text{ for } i = 1, 2, \cdots, q$$

- Assign the test pattern x to class C_l if

$$P(C_l|x) \geq P(C_i|x), \text{ for } i = 1, 2, \cdots, q$$

- In the case of a tie (two or more of the largest-valued posteriors are equal), assign arbitrarily to any one of the corresponding classes. In practice, breaking the tie arbitrarily is the prescription suggested for any ML model.
- In this case, the probability of error is the sum of the posterior probabilities of the remaining $q-1$. We know that the posteriors across all the q classes add up to 1, that is, $\sum_{i=1}^{q} P(C_i|x) = 1$. So, if x is assigned to class C_l, then the probability of error is $1 - P(C_l|x)$.
- In this case also, we have average probability of error as

$$\int_x P(error|x)P(x)dx = \int_x (1 - P(C_l|x))P(x)dx$$

This is the minimum possible because for every x, we are choosing the class that has the largest posterior. So, $P(C_l|x)$ is the largest and $1 - P(C_l|x)$ is minimized for x. So, even in the multi-class case, the Bayes classifier is optimal by being the minimum error rate classifier. It can deal with a mix of both categorical and numerical attributes provided the required probabilities are known.

So, the Bayes classifier is an optimal classifier and is the ideal choice for classification. However, it is used more as a benchmark classifier for theoretical comparisons. In practice, it is difficult to obtain the underlying probability structure. Some of related simplifications that are popular in practice are discussed in the next two sections.

4.6 Parametric and Non-Parametric Schemes for Density Estimation

The Bayes classifier is an optimal classifier. It is versatile in terms of being used in applications involving mixed variables. However, a major bottleneck in its effective use is the assumption that the underlying probability structure is available. We need to have the prior probabilities and the PDF or PMF for each class. We will consider the estimation of prior probabilities.

Example 17: Consider a tweet, and out of 100 people in a community, let 10 from set $\{1, 3, 12, 21, 33, 54, 66, 75, 84, 93\}$ have retweeted while the remaining 90 did not.

It is possible to view the retweeting pattern x as a binary string of length 100, where the i^{th} bit is 1 if the i^{th} person has retweeted for $i = 1, 2, \cdots, 100$, else it is 0.

Let p be the probability of retweet; then the corresponding probability may be captured by $p^{x^i}(1-p)^{1-x^i}$. Note that this quantity selects p if $x^i = 1$ and (1−p) if $x^i = 0$, where x^i is the i^{th} bit of x.

We assume that people retweet independently; then the joint probability is the product of the individual probabilities. So, the joint probability is $p \cdots (1-p) \cdots p \cdots (1-p) = p^{10}(1-p)^{90}$. This is the likelihood of 10 out of 100 people retweeting and the remaining 90 not retweeting. The logarithm of the likelihood is $l(p) = 10logp + 90log(1-p)$. Note that the maximum value of the likelihood is the same as the maximum value of the logarithm of the likelihood as logarithm is a monotonic function.

The derivative of $l(p)$ with respect to p gives us $\frac{10}{p} - \frac{90}{1-p} = 0 \Rightarrow 100p = 10$. So, $p = 0.1 = \frac{10}{100}$; it is actually an estimate of p called the **maximum likelihood estimate (MLE)**. If k out of n people have retweeted, then MLE of p is $p = \frac{k}{n}$. It is called the maximum likelihood estimate because the estimate maximizes the joint probability or likelihood.

4.6.1 Parametric Schemes

Here, we typically assume that the form of the underlying PDF or PMF is known and estimate the parameters involved. In Example 15, we looked at n independent trials of a Bernoulli RV which amounted to the Binomial distribution with $E[k] = np \Rightarrow \hat{p} = \frac{E[k]}{n} = \frac{k}{n}$ (refer to Exercise 10 at the end of this chapter for more details on $E[k] = np$.)

Maximum Likelihood Estimation (MLE)

The maximum likelihood estimate appears to be a simple and intuitively appealing scheme for dealing with estimation of parameters. We will consider the case of a continuous RV.

A quick comparison of the MLE and BE schemes is given below:

- Both MLE and BE are *parametric estimation schemes*. Both of them assume that the functional form of the underlying density is known and the parameters need to be estimated. For example, if we assume that the density is binomial, we need to estimate the value of p that corresponds to the probability of success; the probability of failure is $1 - p$.
- Both of them assume that the training data points are drawn ***independently*** from the unknown density.
- MLE assumes that the parameters are unknown but are deterministic quantities. For example, the estimate of p in n Bernoulli trials (binomial) is $\hat{p} = \frac{k}{n}$, where k is the number of successes out of n trials. However, BE assumes that the unknown parameters are random; for example, in the case of Bernoulli data, we assumed that the unknown parameter p is an RV and it has beta density.
- In the case of normal density, MLE assumes that the unknown mean is deterministic and it estimates mean, $\hat{\mu}$, using the training data only. The estimate is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$, as discussed in an earlier section.
- In the case of BE, for estimating the mean of the normal:
  - The conjugate prior density is normal with mean μ_{in} and variance σ_{in}^2.
  - Because of the use of conjugate prior, the posterior is also normal. If the posterior is normal with mean μ_f and variance σ_f^2, then it is possible to show that

$$\mu_f = \frac{n\sigma_{in}^2}{n\sigma_{in}^2 + \sigma^2} \times \mu_n + \frac{\sigma^2}{n\sigma_{in}^2 + \sigma^2} \times \mu_{in},$$

    where the data is normally distributed with mean μ and variance σ^2 and μ_n is the sample mean or the MLE.
  - Note that the estimate of the mean by the BE scheme coincides again with the estimate of the MLE scheme when $n \to \infty$. Then we have $\mu_f = \mu_n$ as $\frac{\sigma^2}{n\sigma_{in}^2 + \sigma^2} \to 0$ as $n \to \infty$ and $\frac{n\sigma_{in}^2}{n\sigma_{in}^2 + \sigma^2} \to \frac{n\sigma_{in}^2}{n\sigma_{in}^2} = 1$ as $n \to \infty$.
  - Note that BE gives the estimate of the mean as a weighted combination of the MLE value μ_n and the mean of the prior given by μ_{in}. These weights are $\frac{n\sigma_{in}^2}{n\sigma_{in}^2 + \sigma^2}$ and $\frac{\sigma^2}{n\sigma_{in}^2 + \sigma^2}$ and they add up to 1; such a weighted combination is called a **convex combination**.
  - So, BE is a generalized version, of which MLE may be viewed as a special case.

4.7 Class Conditional Independence and Naïve Bayes Classifier

In this section, we examine the difficulties associated with the practical usage of the Bayes classifier and then provide a simplified scheme for estimating the probability structure.

4.7.1 Estimation of the Probability Structure

There are two different schemes for the estimation of parameters: non-parametric schemes and parametric schemes.

Non-Parametric Schemes

Here, the training data is used to directly estimate the PDF. We will consider the non-parametric schemes first to estimate the probability density of a class.

Let n independently drawn training patterns be given from a class. Let each of them be an l-dimensional vector. Let the probability that any one of them falls in a small region, R, in the l-dimensional space be P_R. Let some k out of the n patterns fall in R. The RV here is binomially distributed as out of n independent trials, some k fall in R and the remaining $n-k$ fall outside R. The expected value of k, $E[k] = n \times P_R$ (refer to Exercise 10 at the end of this chapter). Using its expected/average value as an estimate of k, $\hat{k}$, we get $\hat{k} = n \times P_R$. So, $P_R = \frac{\hat{k}}{n}$.

Assume that $k = \hat{k}$; this assumption is valid when the training data size n is large. Then

$$k = n \times P_R \Rightarrow P_R = \frac{k}{n}$$

If the unknown PDF is $p_X(x)$, then P_R is obtained by taking the integral of the PDF in the region R. So, $P_R = \int_R p(X)dx$. But we have $P_R = \frac{k}{n}$. So, from these two we get

$$k = n \times \int_R p_X(x)$$

If we assume that $p_X(x)$ is some constant p, in the region considered, because the region R is very small, then

$$k = n \times p \times V_R,$$

where we get

$$\int_R p_X(x)dx = p \int_R dx = pV_R$$

based on p being constant and V_R being the volume of the region.
So, the estimate of the PDF, p, in a small region is given by

$$p = \frac{1}{n}\frac{k}{V_R}$$

This estimate makes sense when the value n is large and the region R is very small. So, the non-parametric schemes demand very large value for n with the density being constant in R. Hence, they are not widely used in practice.

Parametric Schemes

Here, we assume that the functional form of the PDF is given and we need to estimate the underlying parameters. There are two popular schemes:

- **Maximum Likelihood Estimation (MLE):** Here, likelihood corresponding to the n independently drawn training patterns is maximized to find the estimates of the underlying parameters. The estimate obtained is such that the probability of generating the given patterns from the resulting distribution is maximum.
- **Bayesian Estimation (BE):** In this case, in addition to the assumptions made by MLE, the parameters are assumed to be RVs with known prior distribution based on domain knowledge. The priors are converted into posterior probabilities using the likelihood of the training data. A popular scheme is to use the maximum a posteriori (MAP) estimate.

We start with an example to make the ideas clear.

EXAMPLE 21: Consider the data shown in Table 4.5. There are two classes, C_0 and C_1, and three training patterns from each of the classes. Each pattern is characterized by three binary features. There is a test pattern in the form of pattern 7. We will use MLE and BE to estimate the probabilities using the 6 training patterns and use the Bayes classifier to classify pattern 7.

TABLE 4.5 Data set to illustrate the difficulties associated with the Bayes classifier

Pattern	Feature1	Feature2	Feature3	Class
1	0	0	0	C_0
2	1	0	1	C_1
3	1	0	0	C_0
4	1	1	1	C_1
5	0	1	1	C_1
6	0	1	1	C_0
7	1	0	1	?

Let us consider the Bayes classifier. Note that the prior probabilities are $P(C_0) = P(C_1) =$ using either MLE or BE.

Consider pattern 7. The corresponding posterior probabilities are:

- Using Bayes' rule, we have $P(C_0|Feature1 = 1, Feature2 = 0, Feature3 = 1)$
  $\propto P(C_0) \times P(Feature1 = 1, Feature2 = 0, Feature3 = 1|C_0)$
  $\propto P(C_0) \times P(Feature1 = 1|C_0) \times P(Feature2 = 0|Feature1 = 1, C_0) \times P(Feature3 = 1|Feature1 = 1, Feature2 = 0, C_0)$.
- Similarly, using Bayes' rule, we get $P(C_1|Feature1 = 1, Feature2 = 0, Feature3 = 1) \propto P(C_1) \times P(Feature1 = 1, Feature2 = 0, Feature3 = 1|C_1)$
  $\propto P(C_1) \times P(Feature1 = 1|C_1) \times P(Feature2 = 0|Feature1 = 1, C_1) \times P(Feature3 = 1|Feature1 = 1, Feature2 = 0, C_1)$

1. The MLE estimates are as follows:
   - For class C_0:
     - $P(Feature1 = 1|C_0) = \frac{1}{3}$
     - $P(Feature2 = 0|Feature1 = 1, C_0) = \frac{1}{1} = 1$
     - $P(Feature3 = 1|Feature1 = 1, Feature2 = 0, C_0) = \frac{0}{1} = 0$
   - For class C_1:
     - $P(Feature1 = 1|C_1) = \frac{2}{3}$
     - $P(Feature2 = 0|Feature1 = 1, C_1) = \frac{1}{2}$
     - $P(Feature3 = 1|Feature1 = 1, Feature2 = 0, C_1) = \frac{1}{1} = 1$

   So, using the MLE estimates, $P(C_0|Feature1 = 1, Feature2 = 0, Feature3 = 1) \propto \frac{1}{2} \times \frac{1}{3} \times 1 \times 0 = 0$.
   Using the MLE estimates for class C_1, we have $P(C_1|Feature1 = 1, Feature2 = 0, Feature3 = 1) \propto \frac{1}{2} \times \frac{2}{3} \times \frac{1}{2} \times 1 = \frac{1}{6}$.
   So, $P(C_1|Feature1 = 1, Feature2 = 0, Feature3 = 1)(= \frac{1}{6} > P(C_0|Feature1 = 1, Feature2 = 0, Feature3 = 1)(= 0)$.
   So, pattern 7 is assigned to C_1 by employing the estimates obtained using the MLE scheme.

2. The BE estimates are as follows:
   - For Class C_0:
     - $P(Feature1 = 1|C_0) = \frac{1+1}{3+2} = \frac{2}{5}$
     - $P(Feature2 = 0|Feature1 = 1, C_0) = \frac{1+1}{1+2} = \frac{2}{3}$
     - $P(Feature3 = 1|Feature1 = 1, Feature2 = 0, C_0) = \frac{0+1}{1+2} = \frac{1}{3}$
   - For Class C_1:
     - $P(Feature1 = 1|C_1) = \frac{2+1}{3+2} = \frac{3}{5}$
     - $P(Feature2 = 0|Feature1 = 1, C_1) = \frac{1+1}{2+2} = \frac{1}{2}$
     - $P(Feature3 = 1|Feature1 = 1, Feature2 = 0, C_1) = \frac{1+1}{1+2} = \frac{2}{3}$

   So, $P(C_0|Feature1 = 1, Feature2 = 0, Feature3 = 1) \propto \frac{1}{2} \times \frac{2}{5} \times \frac{2}{3} \times \frac{1}{3} = \frac{2}{45}$.
   Similarly, $P(C_1|Feature1 = 1, Feature2 = 0, Feature3 = 1) \propto \frac{1}{2} \times \frac{3}{5} \times \frac{1}{2} \times \frac{2}{3} = \frac{1}{10}$.
   So, $P(C_1|Feature1 = 1, Feature2 = 0, Feature3 = 1) > P(C_0|Feature1 = 1, Feature2 = 0, Feature3 = 1)$.
   So pattern 7 is assigned to C_1 by employing the estimates obtained using the BE scheme.

Some observations based on Example 21:

- Both the MLE and BE schemes have assigned pattern 7 to the same class, that is, C_1. This is not the case in general. We will examine it in Exercise 12 at the end of this chapter.
- It is possible that the MLE scheme estimates zero posterior probabilities for two or more classes, leading to a difficulty in taking a meaningful decision. This problem will not be encountered when we use BE. This property also will be examined in Exercise 12 at the end of this chapter.
- On large data sets, that is, when $n \to \infty$, the BE scheme gives the same estimate as the MLE scheme. So, the recommendation is to use the BE scheme for estimation when the training data is small by integrating the domain knowledge in the form of a suitable prior.
- On larger sized training data, it is good to use the MLE estimate as it depends solely on the data.

4.7.2 Naïve Bayes Classifier (NBC)

We have discussed the difficulties associated with the use of the Bayes classifier in practice. One simplification that is popular is based on *class-conditional independence*. The resulting classifier is called the naïve Bayes classifier as it is a Bayes classifier with some simplification. It may be explained using an example.

EXAMPLE 22: Consider the data used in Example 21 and Table 4.5. Let us again consider the posterior probabilities for pattern 7. They are:

$$P(C_0|pattern\ 7) = \frac{P(Feature1=1, Feature2=0, Feature3=1|C_0) \times P(C_0)}{P(Feature1=1, Feature2=0, Feature3=1)} \text{ and}$$

$$P(C_1|pattern\ 7) = \frac{P(Feature1=1, Feature2=0, Feature3=1|C_1) \times P(C_1)}{P(Feature1=1, Feature2=0, Feature3=1)}$$

Note that both the posteriors have the same denominator. So, instead of comparing the posteriors, we can compare their numerators for the sake of simplicity as was done in the previous subsection.

We have $P(C_0) = P(C_1) = 0.5$. We need to compute the likelihood values for the two classes. We can simplify the computation using class-conditional independence as follows:

- For Class C_0, $P(Feature1 = 1, Feature2 = 0, Feature3 = 1|C_0)$
  $= P(Feature1 = 1|C_0) \times P(Feature2 = 0|C_0) \times P(Feature3 = 1|C_0)$
- Similarly, for C_1, $P(Feature1 = 1, Feature2 = 0, Feature3 = 1|C_1)$
  $= P(Feature1 = 1|C_1) \times P(Feature2 = 0|C_1) \times P(Feature3 = 1|C_1)$

What we have done is to write the probability of a conjunction conditioned on Class C_0 or C_1 as the product of probabilities of individual conjuncts that are conditioned on the class. This is **class-conditional independence.**

The required probabilities may be calculated using the MLE scheme for Class C_0 as:

- $P(Feature1 = 1|C_0) = \frac{1}{3}$.
- $P(Feature2 = 0|C_0) = \frac{2}{3}$.
- $P(Feature3 = 1|C_0) = \frac{1}{3}$.

  So, $P(C_0|pattern\ 7) \propto \frac{1}{2} \times \frac{1}{3} \times \frac{2}{3} \times \frac{1}{3} = \frac{1}{27}$

Similarly, for C_1 we have

- $P(Feature1 = 1|C_1) = \frac{2}{3}$.
- $P(Feature2 = 0|C_1) = \frac{1}{3}$.
- $P(Feature3 = 1|C_1) = \frac{3}{3} = 1$.

  So, $P(C_1|pattern\ 7) \propto \frac{1}{2} \times \frac{2}{3} \times \frac{1}{3} \times 1 = \frac{1}{9}$

So, we assign pattern 7 to C_1 as $P(C_1|pattern\ 7) > P(C_0|pattern\ 7)$

Some general observations from Example 22:

- In this example, NBC has made the same decision as the Bayes classifier. However, they can give different results in general.
- NBC employs class-conditional independence. It is given in general as

$$P(f_1 = v_1, f_2 = v_2, \cdots, f_l = v_l|C) = P(f_1 = v_1|C) \times P(f_2 = v_2|C) \times \cdots \times P(f_l = v_l|C)$$

  where C is the class and $f_i = v_i$ means feature i (f_i) will have value v_i.
- The decisions of NBC coincide with that of the Bayes classifier if class-conditional independence holds.
- NBC has simplified the computation and even the MLE scheme estimates did not create a problem here as the estimates of these simpler probabilities are non-zero.
- We have used the MLE scheme to estimate the probabilities. It is possible to use the BE scheme to ensure that we do not have zero-valued estimates. This is left to the reader as an exercise.

Summary

In this chapter, we examined ML models based on Bayes' rule. Some important issues discussed are as follows:

- Bayes' rule plays an important role in understanding the Bayes classifier.
- The Bayes classifier is an optimal classifier. It minimizes the probability of error.
- Akin to the classifiers based on DTs, the Bayes classifier can be used when the data set has mixed type features, that is, categorical and numerical. However, we require the underlying probability structure to use the Bayes classifier effectively.
- It is possible to estimate the probability structure using training data with the help of the maximum likelihood approach or the Bayesian scheme.

- Estimation of the probability structure can be simplified by assuming class-conditional independence.
- NBC makes use of class-conditional independence and is a popular choice when the data set has mixed type features.
- NBC is a linear classifier and hence is a good candidate for providing explanation to the users and domain experts.

EXERCISES

1. Consider the discussion immediately after Example 1. Show that KNN is equivalent to prior probability-based classification, when $k = n$ and the probabilities are estimated based on ratio of frequencies.
2. Consider tossing a coin twice. What is the sample space? Find the probability that the first toss results in a tail. What is the probability that the second toss results in a head?
3. Show that $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.
4. If events A and B are independent, then show that A^c and B are independent.
5. Suppose a fair coin is tossed three times. Let A be the event that we get two heads in these three tosses. Let B be the event that the first toss shows up heads. Obtain
   a. $P(A)$ and $P(B)$
   b. $P(A|B)$ and $P(B|A)$
6. Consider the joint probability function $P_{X,Y}(x, y)$ given in Table 4.2. Obtain the marginal probability values of $P_Y(y)$ for all possible values of y.
7. Consider the discussion on binomial RV at the end of Section 4.3.3. The probability of k tails out of n tosses is given by $P_X(k)$. Show that
   a. $P_X(k) \geq 0$ for $0 \leq k \leq n$, and
   b. $\sum_{k=0}^{n} P_X(k) = 1$.
8. Consider the discussion before and in Example 13 of computing the expected value of a function of an RV. Let X be an RV with its expectation $E[X]$. If $h(X) = a \times X + b$, where a and b are some constants, find $E[h(X)]$.
9. Consider an RV X that is uniformly distributed in the range (0,1). Plot its PDF, $f_X(x)$, and CDF, $F_X(x)$.
10. Given that the mean of a Bernoulli-distributed RV is p and variance is $p(1-p)$ (refer Example 14), show that the mean of the binomial RV is np and its variance is $np(1-p)$ using the fact that binomial RV corresponds to n independent trails of the Bernoulli RV. If a coin is tossed n times (independently), getting k heads is binomially distributed. So, mean of binomial RV is $E[k]$ and variance is $E[(k - E[k])^2]$.
11. Show that if two classes, C_1 and C_2, are normally distributed with equal priors and covariance matrices being equal to Σ, then it is optimal to assign x to C_1 if
$$(x - \mu_1)^t \Sigma^{-1} (x - \mu_1) < (x - \mu_2) \Sigma^{-1} (x - \mu_2).$$
12. Consider the discussion in Example 22. Classify pattern 7 using NBC and estimate the probabilities using the BE scheme. Assume that prior probabilities are equal to $\frac{1}{2}$ for both the classes.

13. Consider the training data used in Example 21 and shown in Table 4.5. Classify the test pattern for which $Feature1 = 1$, $Feature2 = 1$ and $Feature3 = 0$ using the Bayes classifier.
    a. Use the MLE scheme for estimating the probabilities. Is there any problem?
    b. Use the BE scheme for estimating the probabilities.
14. Solve Q13 using NBC instead of the Bayes classifier. Use the MLE and BE schemes to estimate the probabilities.

Practical Exercise

1. Download the Olivetti Face data set. There are 40 classes (corresponding to 40 people), each class having 10 faces of the individual; so there are a total of 400 images. Here, each face is viewed as an image of size 64×64 ($= 4096$) pixels; each pixel has values 0 to 255 which are ultimately converted into floating numbers in the range [0,1]. Visit `https://scikit-learn.org/0.19/datasets/olivetti\_faces.html` for more details.
   Split the data sets into train and test parts. Perform this splitting randomly 10 times and report the average accuracy. You may vary the test and train data set sets. Use NBC to classify the test data set. Obtain the accuracy on the test data.

Bibliography

- Murty, MN and Susheela Devi, V. 2015 *Introduction to Pattern Recognition and Machine Learning*, World Scientific Publishing Co. Pte. Ltd.: Singapore.
- Duda, RO, Hart, PE and Stork, DG. 2007. *Pattern Classification*, New York: John Wiley & Sons.
- Han, J, Kamber, M and Pei, J. 2012. *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers: Waltham.
- Witten, IH, Frank, E and Hall, MA. 2011. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, Morgan Kaufmann Publishers, Burlington.
- Tan, P-N, Steinbach, M, Karpatne, A and Kumar, V. 2018. *Introduction to Data Mining, Second Edition*, Pearson.