Chapter 21

Probabilistic Classification

21.1 Bayes Classifier

Let the training dataset $\mathbf{D}$ consist of n points $\mathbf{x}_i$ in a d-dimensional space, and let y_i denote the class for each point, with $y_i \in \{c_1, c_2, \cdots, c_k\}$. The Bayes classifier makes uses of the Bayes theorem to predict the class for a new test instance, $\mathbf{x}$. It estimates the posterior probability $P(c_i|\mathbf{x})$ for each class c_i, and chooses the class that has the largest probability. That is, the predicted label $\hat{y}$ for $\mathbf{x}$ is given as

$$\hat{y} = \arg\max_i \{P(c_i|\mathbf{x})\} \tag{21.1}$$

The Bayes theorem allows us to invert the posterior probability in terms of the likelihood and prior probability, as follows

$$P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i) \cdot P(c_i)}{P(\mathbf{x})} \tag{21.2}$$

Here $P(\mathbf{x}|c_i)$ is the *likelihood*, defined as the probability of observing $\mathbf{x}$, assuming that the true class is c_i. $P(c_i)$ is the *prior probability* of class c_i, and $P(\mathbf{x})$ is the probability of observing $\mathbf{x}$ from any of the k classes, given as

$$P(\mathbf{x}) = \sum_{j=1}^{k} P(\mathbf{x}|c_j) \cdot P(c_j) \tag{21.3}$$

Since $P(\mathbf{x})$ is fixed for a given point, the Bayes rule in (21.1) can be rewritten as

$$\begin{aligned} \hat{y} &= \arg\max_i \{P(c_i|\mathbf{x})\} = \arg\max_i \left\{ \frac{P(\mathbf{x}|c_i)P(c_i)}{P(\mathbf{x})} \right\} \\ &= \arg\max_i \big\{ P(\mathbf{x}|c_i)P(c_i) \big\} \end{aligned} \tag{21.4}$$

In other words, the predicted class essentially depends on the likelihood of that class, taking its prior probability into account.

21.1.1 Prior Probability

To classify points, we have to estimate the likelihood and prior probabilities directly from the training dataset $\mathbf{D}$. Let $\mathbf{D}_i$ denote the subset of points in $\mathbf{D}$ that are labeled with class c_i, given as

$$\mathbf{D}_i = \{\mathbf{x}_j \in \mathbf{D} \mid \mathbf{x}_j \text{ has label } y_j = c_i\} \tag{21.5}$$

Let the size of the dataset $\mathbf{D}$ be given as $|\mathbf{D}| = n$, and let the size of each class-specific subset $\mathbf{D}_i$ be given as $|\mathbf{D}_i| = n_i$. The prior probability for class c_i can be estimated as follows

$$\hat{P}(c_i) = \frac{n_i}{n} \tag{21.6}$$

21.1.2 Likelihood

To estimate the likelihood $P(\mathbf{x}|c_i)$, we have to estimate the joint probability of $\mathbf{x}$ across all the d dimensions, i.e., we have to estimate $P(\mathbf{x} = (x_1, x_2, \cdots, x_d)|c_i)$.

Numeric Attributes Assuming all dimensions are numeric, we can estimate the joint probability of $\mathbf{x}$ via either a non-parametric or a parametric approach.

In the non-parametric approach we compute the empirical joint probability density function directly from the data sample $\mathbf{D}_i$ for class c_i. This can be done using the kernel density estimation methods from Chapter 15.

In the parametric approach we assume that each class c_i is normally distributed around some mean $\boldsymbol{\mu}_i$, with a corresponding covariance matrix $\boldsymbol{\Sigma}_i$. The likelihood for class c_i for a test point $\mathbf{x}$ is given as

$$f_i(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\boldsymbol{\Sigma}_i|}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)}{2}\right\} \tag{21.7}$$

Since c_i is characterized by a continuous distribution, the probability of any given point must be zero, i.e, $P(\mathbf{x}|c_i) = 0$. However, we can compute the likelihood by considering a small interval $\epsilon > 0$ around $\mathbf{x}$

$$P(\mathbf{x}|c_i) = 2\epsilon \cdot f_i(\mathbf{x})$$

The posterior probability is then given as

$$P(c_i|\mathbf{x}) = \frac{2\epsilon \cdot f_i(\mathbf{x}) P(c_i)}{\sum_{i=1}^k 2\epsilon \cdot f_i(\mathbf{x}) P(c_i)} = \frac{f_i(\mathbf{x}) P(c_i)}{\sum_{i=1}^k f_i(\mathbf{x}) P(c_i)} \tag{21.8}$$

Further, since $\sum_{i=1}^k f_i(\mathbf{x}) P(c_i)$ remains fixed for $\mathbf{x}$, we can predict the class for $\mathbf{x}$ by modifying (21.4) as follows

$$\hat{y} = \arg\max_i \Big\{ f_i(\mathbf{x}) P(c_i) \Big\}$$

To classify a numeric test point $\mathbf{x}$, the (full) Bayes classifier estimates the parameters via the sample mean and sample covariance matrix. The sample mean for the class c_i can be estimated as (2.44)

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in \mathbf{D}_i} \mathbf{x}_j \tag{21.9}$$

The sample covariance matrix for each class can be estimated via (2.50) as follows

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i} \mathbf{Z}_i^T \mathbf{Z}_i \tag{21.10}$$

where $\mathbf{Z}_i$ is the centered data matrix for class c_i given as $\mathbf{Z}_i = \mathbf{D}_i - \mathbf{1} \cdot \hat{\boldsymbol{\mu}}_i^T$. These values can be used to estimate the likelihood (21.7), $\hat{f}_i(\mathbf{x}) = \hat{f}(\mathbf{x}|\hat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i)$.

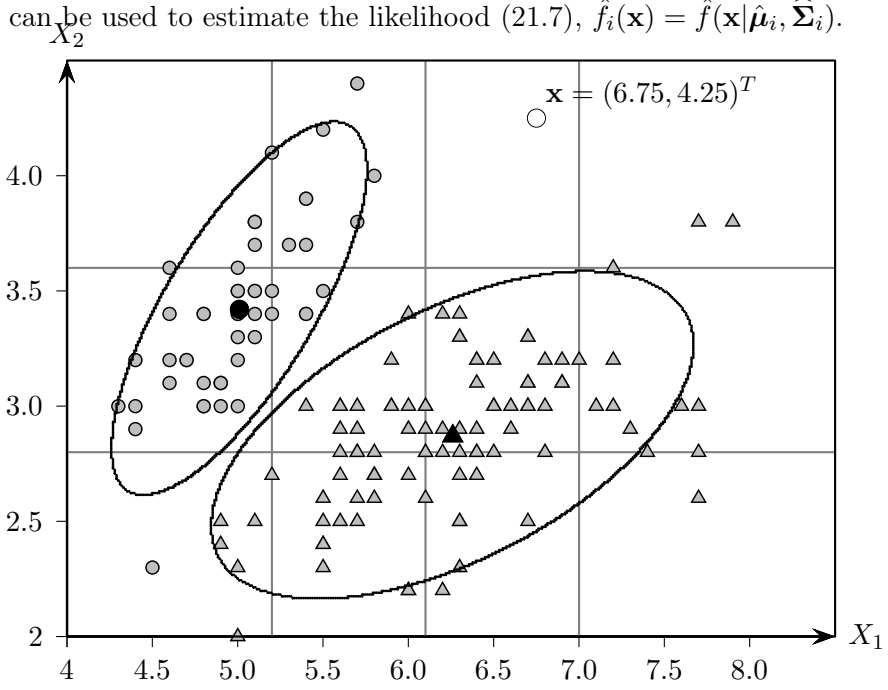

Figure 21.1: Iris Data: `sepal length versus sepal width`

Example 21.1: Consider the two-dimensional Iris data (for `sepal length` and `sepal width`) shown in Figure 21.1. Class c_1, corresponding to `iris-setosa` (shown as circles), has $n_1 = 50$ points, whereas the other class c_2 (shown as triangles) has $n_2 = 100$ points. Thus their prior probabilities are

$$\hat{P}(c_1) = \frac{n_1}{n} = \frac{50}{150} = 0.33$$
$$\hat{P}(c_2) = \frac{n_2}{n} = \frac{100}{150} = 0.67$$

The means for c_1 and c_2 (shown as black circle and triangle) are given as

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} 5.01 & 3.42 \end{pmatrix}^T$$
$$\hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} 6.26 & 2.87 \end{pmatrix}^T$$

Finally, their covariance matrices are as follows

$$\widehat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.122 & 0.098 \\ 0.098 & 0.142 \end{pmatrix}$$
$$\widehat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.435 & 0.121 \\ 0.121 & 0.110 \end{pmatrix}$$

Figure 21.1 shows the contour or level curve (corresponding to 1% of the peak density) of the multivariate normal distribution for both classes.

Let $\mathbf{x} = (6.75, 4.25)^T$ be a test point. The posterior probabilities for c_1 and c_2 can be computed using (21.8)

$$\hat{P}(c_1|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}}_1)\hat{P}(c_1) = (4.914 \times 10^{-7}) \times 0.33 = 1.622 \times 10^{-7}$$
$$\hat{P}(c_2|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}_2)\hat{P}(c_2) = (2.589 \times 10^{-5}) \times 0.67 = 1.735 \times 10^{-5}$$

Since $\hat{P}(c_2|\mathbf{x}) > \hat{P}(c_1|\mathbf{x})$ we predict the class as $\hat{y} = c_2$.

Categorical Attributes If the attributes are categorical, the likelihood is computed via the approach in Chapter 3. We model each categorical attribute X_j as a vector random variable $\mathbf{V}_j$ that takes on m_j distinct vector values $\mathbf{e}_{j1}, \mathbf{e}_{j2}, \ldots, \mathbf{e}_{jm_j}$, where $\mathbf{e}_{jr}$ corresponds to the r-th value or symbol $a_{jr} \in dom(X_j)$. The categorical point $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$, with $x_j \in dom(X_j)$ is represented as the point $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d)$, where $\mathbf{v}_j = \mathbf{e}_{jr_j}$. The joint probability of the categorical point $\mathbf{v}$ is obtained from the joint PMF for the vector random variable $\mathbf{V} = \begin{pmatrix} \mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_d \end{pmatrix}^T$, given as

$$P(\mathbf{x}|c_i) = f(\mathbf{v}|c_i) = f(\mathbf{V}_1 = \mathbf{e}_{1j_1}, \ldots, \mathbf{V}_d = \mathbf{e}_{dj_d}) \tag{21.11}$$

We can estimate the joint PMF via the empirical joint PMF given as

$$\hat{f}(\mathbf{v}|c_i) = \frac{n_i(\mathbf{v})}{n_i} \tag{21.12}$$

where $n_i(\mathbf{v})$ is the number of times the value $\mathbf{v}$ occurs in class c_i. Unfortunately, if the probability mass at the point $\mathbf{v}$ is zero for one or both classes, it would lead to a zero value for the posterior probability. To avoid zero probabilities, it is customary

to introduce a small prior probability for all the possible values of the vector random variable $\mathbf{V}$. One simple approach is to assume a *pseudo-count* of 1, leading to a prior probability of $1/n_i$ for each possible value of $\mathbf{V}$ for the class c_i. The adjusted probability mass at $\mathbf{v}$ is given as

$$\hat{f}(\mathbf{v}|c_i) = \frac{n_i(\mathbf{v}) + 1}{n_i + \prod_{j=1}^{k} m_j} \tag{21.13}$$

where $m_j = |dom(X_j)|$, and $\prod_{j=1}^{k} m_j$ gives the total number of distinct values of $\mathbf{V}$.

bins	domain
$[4.3, 5.2]$	Very Short (a_{11})
$(5.2, 6.1]$	Short (a_{12})
$(6.1, 7.0]$	Long (a_{13})
$(7.0, 7.9]$	Very Long (a_{14})

(a) Discretized `sepal length`

bins	domain
$[2.0, 2.8]$	Short (a_{21})
$(2.8, 3.6]$	Medium (a_{22})
$(3.6, 4.4]$	Long (a_{23})

(b) Discretized `sepal width`

Table 21.1: Discretized `sepal length` and `sepal width` Attributes

	Class: c_1	X_2			$\hat{f}_{X_1}$
		`Short` ($\mathbf{e}_{21}$)	`Medium` ($\mathbf{e}_{22}$)	`Long` ($\mathbf{e}_{23}$)	
X_1	`Very Short` ($\mathbf{e}_{11}$)	1/50	33/50	5/50	39/50
	`Short` ($\mathbf{e}_{12}$)	0	3/50	8/50	13/50
	`Long` ($\mathbf{e}_{13}$)	0	0	0	0
	`Very Long` ($\mathbf{e}_{14}$)	0	0	0	0
	$\hat{f}_{X_2}$	1/50	36/50	13/50	

	Class: c_2	X_2			$\hat{f}_{X_1}$
		`Short` ($\mathbf{e}_{21}$)	`Medium` ($\mathbf{e}_{22}$)	`Long` ($\mathbf{e}_{23}$)	
X_1	`Very Short` ($\mathbf{e}_{11}$)	6/100	0	0	6/100
	`Short` ($\mathbf{e}_{12}$)	24/100	15/100	0	39/100
	`Long` ($\mathbf{e}_{13}$)	13/100	30/100	0	43/100
	`Very Long` ($\mathbf{e}_{14}$)	3/100	7/100	2/100	12/100
	$\hat{f}_{X_2}$	46/100	52/100	2/100	

Table 21.2: Class-specific Empirical (Joint) Probability Mass Function

Example 21.2: Assume that the `sepal length` and `sepal width` attributes in the Iris dataset have been discretized as shown in Table 21.1a and Table 21.1b.

These intervals are also illustrated in Figure 21.1 via the gray grid lines. Table 21.2 shows the empirical joint PMF for both the classes.

Consider a test point $\mathbf{x} = (5.3, 3.0)$, which yields the categorical point (`Short, Medium`), represented as $\mathbf{v} = \begin{pmatrix}\mathbf{e}_{12} & \mathbf{e}_{22}\end{pmatrix}$. The likelihood and posterior probability for each class is given as

$$\hat{P}(\mathbf{x}|c_1) = \hat{P}(\mathbf{v}|c_1) = 3/50 = 0.06$$
$$\hat{P}(\mathbf{x}|c_2) = \hat{P}(\mathbf{v}|c_2) = 15/100 = 0.15$$
$$P(c_1|\mathbf{x}) \propto 0.06 \times 0.33 = 0.0198$$
$$P(c_2|\mathbf{x}) \propto 0.15 \times 0.67 = 0.1005$$

In this case the predicted class is $\hat{y} = c_2$.

On the other hand, the test point $\mathbf{x} = (6.75, 4.25)$ corresponding to the categorical point (`Long, Long`), is represented as $\mathbf{v} = \begin{pmatrix}\mathbf{e}_{13} & \mathbf{e}_{23}\end{pmatrix}$. Unfortunately the probability mass at this value is zero for both classes. Adjusting the PMF via the pseudo-counts method given in (21.13), the likelihood and prior probability for each class for the point $\mathbf{v} = \begin{pmatrix}\mathbf{e}_{13} & \mathbf{e}_{23}\end{pmatrix}$ is given as

$$\hat{P}(\mathbf{x}|c_1) = \hat{P}(\mathbf{v}|c_1) = \frac{1}{50+12} = 1.61 \times 10^{-2}$$
$$\hat{P}(\mathbf{x}|c_2) = \hat{P}(\mathbf{v}|c_2) = \frac{1}{100+12} = 8.93 \times 10^{-3}$$
$$\hat{P}(c_1|\mathbf{x}) \propto (1.61 \times 10^{-2}) \times 0.33 = 5.32 \times 10^{-3}$$
$$\hat{P}(c_2|\mathbf{x}) \propto (8.93 \times 10^{-3}) \times 0.67 = 5.98 \times 10^{-3}$$

Thus the predicted class is $\hat{y} = c_2$.

Challenges The main problem with the Bayes classifier is the lack of enough data to reliably estimate the joint probability density or mass function, especially with increasing dimensionality d. For instance, for numeric attributes we have to estimate $O(d^2)$ covariances, and as the dimensionality increases, this requires us to estimate too many parameters. For categorical attributes we have to estimate the joint probability for all the possible values of $\mathbf{v}$, given as $\prod_i \text{dom}(X_i)$. Even if each categorical attribute has only two values, we need to estimate the probability for 2^d values. However, since there can be at most n distinct values for $\mathbf{v}$, most of the counts will be zero.

21.2 Naïve Bayes Classifier

We saw above that the (full) Bayes approach is fraught with estimation related problems, especially with large number of dimensions. The naïve Bayes approach makes the "naïve" assumption that all the attributes are independent. This leads to a much simpler, though surprisingly effective classifier in practice. The independence assumption immediately implies that the likelihood can be decomposed into a product of dimension-wise probabilities

$$P(\mathbf{x}|c_i) = P(x_1, x_2, \cdots, x_d|c_i) = \prod_{j=1}^{d} P(x_j|c_i) \tag{21.14}$$

Numeric Attributes For numeric attributes we make the default assumption that each of them is normally distributed for each class c_i. Let μ_{ij} and σ_{ij}^2 denote the mean and variance for attribute X_j in class c_i. The likelihood for class c_i, for dimension j is given as

$$P(x_j|c_i) = f(x_j|\mu_{ij}, \sigma_{ij}^2) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}\right\} \tag{21.15}$$

Incidentally, the naïve assumption corresponds to setting all the covariances to zero in $\mathbf{\Sigma}_i$

$$\mathbf{\Sigma}_i = \begin{pmatrix} \sigma_{i1}^2 & 0 & \ldots & 0 \\ 0 & \sigma_{i2}^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \ldots & \sigma_{id}^2 \end{pmatrix} \tag{21.16}$$

This yields

$$|\mathbf{\Sigma}_i| = \det(\mathbf{\Sigma}_i) = \sigma_{i1}^2 \sigma_{i2}^2 \cdots \sigma_{id}^2 = \prod_{j=1}^{d} \sigma_{ij}^2 \tag{21.17}$$

Also, we have

$$\mathbf{\Sigma}_i^{-1} = \begin{pmatrix} \frac{1}{\sigma_{i1}^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sigma_{i2}^2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \ldots & \frac{1}{\sigma_{id}^2} \end{pmatrix} \tag{21.18}$$

assuming that $\sigma_{ij}^2 \neq 0$ for all j. Finally,

$$(\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) = \sum_{j=1}^{d} \frac{(x_j - \mu_{ij})^2}{\sigma_{ij}^2} \tag{21.19}$$

Plugging these into (21.7) gives us

$$\begin{aligned} P(\mathbf{x}|c_i) &= \frac{1}{(\sqrt{2\pi})^d \sqrt{\prod_{j=1}^{d} \sigma_{ij}^2}} \exp\left\{ -\sum_{j=1}^{d} \frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2} \right\} \\ &= \prod_{j=1}^{d} \left(\frac{1}{\sqrt{2\pi}\ \sigma_{ij}} \exp\left\{ -\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2} \right\} \right) \\ &= \prod_{j=1}^{d} P(x_j|c_i) \end{aligned} \tag{21.20}$$

which is equivalent to (21.14). In other words, the joint probability has been decomposed into a product of the probability along each dimension, as required by the independence assumption.

The naïve Bayes classifier uses the sample mean $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}_{i1}, \ldots, \hat{\mu}_{id})^T$ and a *diagonal* sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_i = diag(\sigma_{i1}^2, \ldots, \sigma_{id}^2)$ for each class c_i. Thus, in total $2d$ parameters have to be estimated corresponding to the sample mean and sample variance along each dimension X_j.

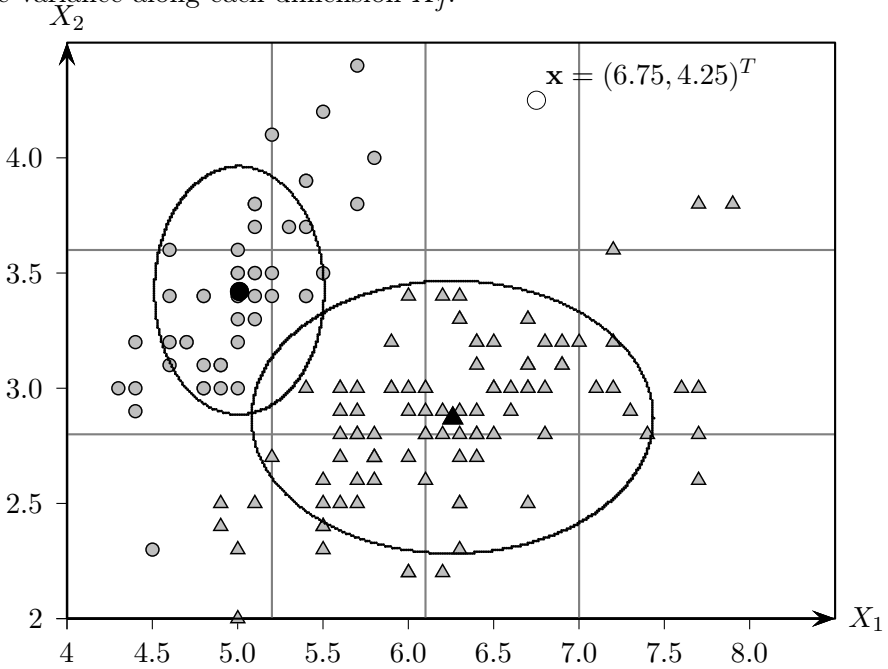

Figure 21.2: Naïve Bayes: `sepal length versus sepal width`

Example 21.3: Consider Example 21.1. In the naïve Bayes approach the prior probabilities $\hat{P}(c_i)$ and means $\hat{\boldsymbol{\mu}}_i$ remain unchanged. The key difference is that the covariance matrices are assumed to be diagonal, as follows

$$\widehat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.122 & 0 \\ 0 & 0.142 \end{pmatrix}$$
$$\widehat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 0.435 & 0 \\ 0 & 0.110 \end{pmatrix}$$

Figure 21.2 shows the contour or level curve (corresponding to 1% of the peak density) of the multivariate normal distribution for both classes. One can see that the diagonal assumption corresponds to the case where the contours are axis-parallel ellipses.

For the test point $\mathbf{x} = (6.75, 4.25)^T$, the posterior probabilities for c_1 and c_2 are as follows

$$\hat{P}(c_1|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}}_1)\hat{P}(c_1) = (3.99 \times 10^{-7}) \times 0.33 = 1.319 \times 10^{-7}$$
$$\hat{P}(c_2|\mathbf{x}) \propto \hat{f}(\mathbf{x}|\hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}_2)\hat{P}(c_2) = (9.597 \times 10^{-5}) \times 0.67 = 6.430 \times 10^{-5}$$

Since $\hat{P}(c_2|\mathbf{x}) > \hat{P}(c_1|\mathbf{x})$ we predict the class as $\hat{y} = c_2$.

Categorical Attributes The independence assumption leads to a univariate modeling of each categorical attribute. For attribute X_j with $dom(X_j) = \{a_{j1}, \ldots, a_{jm_j}\}$, we define a vector random variable $\mathbf{V}_j$ that takes on values $\{\mathbf{e}_{j1}, \ldots, \mathbf{e}_{jm_j}\}$. The joint probability in (21.11) can be rewritten as

$$P(\mathbf{v}|c_i) = \prod_{j=1}^{d} P(\mathbf{v}_j|c_i) \tag{21.21}$$

where $P(\mathbf{v}_j|c_i) = f(\mathbf{v}_j|c_i)$ is the probability mass function for $\mathbf{V}_j$, which can be estimated as follows

$$\hat{f}(\mathbf{v}_j|c_i) = \frac{n_i(\mathbf{v}_j)}{n_i} \tag{21.22}$$

where $n_i(\mathbf{v}_j)$ is the frequency of $\mathbf{v}_j$ for the j-th categorical attribute in class c_i. As in the full Bayes case, if the count is zero, we can use the pseudo-count method to obtain a prior probability. The new estimates with pseudo-counts are given as

$$\hat{f}(\mathbf{v}_j|c_i) = \frac{n_i(\mathbf{v}_j) + 1}{n_i + m_j} \tag{21.23}$$

where $m_j = |dom(X_j)|$.

Example 21.4: Continuing Example 21.2, the class-specific PMF for each discretized attribute is shown in Table 21.2.

The test point $\mathbf{x} = (6.75, 4.25)$, corresponding to (`Long, Long`) or $\mathbf{v} = (\mathbf{e}_{13}, \mathbf{e}_{23})$, is classified as follows

$$\hat{P}(\mathbf{v}|c_1) = \hat{P}(\mathbf{e}_{13}|c_1) \cdot \hat{P}(\mathbf{e}_{23}|c_1) = \left(\frac{1}{50+4}\right) \cdot \left(\frac{13}{50}\right) = 4.815 \times 10^{-3}$$

$$\hat{P}(\mathbf{v}|c_2) = \hat{P}(\mathbf{e}_{13}|c_2) \cdot \hat{P}(\mathbf{e}_{23}|c_2) = \left(\frac{43}{100}\right) \cdot \left(\frac{2}{100}\right) = 8.6 \times 10^{-3}$$

$$\hat{P}(c_1|\mathbf{v}) \propto (4.815 \times 10^{-3}) \times 0.33 = 1.589 \times 10^{-3}$$

$$\hat{P}(c_2|\mathbf{v}) \propto (8.6 \times 10^{-3}) \times 0.67 = 3.226 \times 10^{-3}$$

Thus the predicted class is $\hat{y} = c_2$.