16 Naive Bayes Classification: Predicting with Probabilities

16.1 Introduction to Naive Bayes: Simple but Powerful

16.1.1 The Idea Behind "Naive"

The Naive Bayes classifier is a probabilistic method for assigning labels to samples using Bayes' theorem. It works by estimating, from training data, how likely each class is a priori and how likely the observed features are under each class.

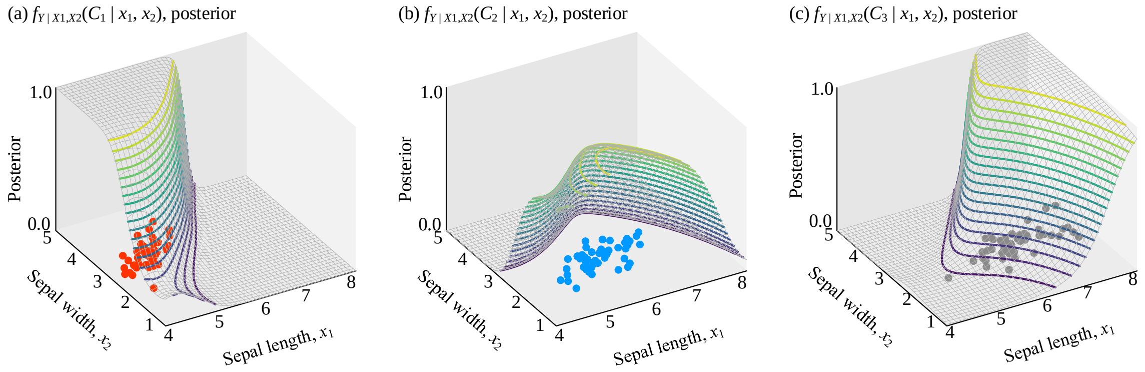
The method is called "naive'' because it assumes the features are conditionally independent given the class — a strong but simplifying assumption that makes the computations straightforward. In practice this assumption often works well enough, especially when there are many features or when you need a fast, scalable baseline.

16.1.2 Posterior Probabilities and Membership Scores

To classify a new example, Naive Bayes computes the posterior probability for each class: intuitively, the probability that the sample belongs to that class given its observed features. These posterior probabilities are sometimes referred to as membership scores because they measure how strongly the point belongs to each class. The classifier then assigns the sample to the class with the largest posterior probability.

Behind the scenes this calculation uses Bayes' theorem: posterior $\propto$ likelihood × prior, so the algorithm combines how well the features fit a class (the likelihood) with how common the class is in the data (the prior). A useful way to understand the method is visually. For a two- or three-feature problem you can plot the posterior as a surface or contour plot over feature space.

For example, Figure 1 (a) shows the posterior surface for class C_1 (setosa), while Figure 1 (b) and Figure 1 (c) show the corresponding surfaces for C_2 (versicolor) and C_3 (virginica).


(a) $f_{Y|X1,X2}(C_1 | x_1, x_2)$, posterior

(b) $f_{Y|X1,X2}(C_2 | x_1, x_2)$, posterior

(c) $f_{Y|X1,X2}(C_3 | x_1, x_2)$, posterior

Figure 1. Posterior probability surfaces and contour plots: (a) $f_{Y|X1,X2}(C_1 | x_1, x_2)$ for setosa, (b) $f_{Y|X1,X2}(C_2 | x_1, x_2)$, and (c) $f_{Y|X1,X2}(C_3 | x_1, x_2)$

Each surface takes values between 0 and 1 and highlights the regions where a class is most probable. At any location in feature space, the class whose posterior surface has the highest value is the predicted label. This visualization highlights an important point: Naive Bayes does not directly draw a decision boundary first; rather, it estimates class probabilities everywhere and lets those probabilities determine the boundary.

That makes the classifier easy to interpret and lets you see how different classes distribute across the feature space. Common practical variants of Naive Bayes model the class-conditional likelihood differently depending on the data type: Gaussian Naive Bayes assumes continuous features follow class-specific normal distributions, Multinomial Naive Bayes models counts (useful for text), and Bernoulli Naive Bayes handles binary features. Each choice affects how the posterior surfaces look and how well the classifier fits a particular problem.

16.2 Estimating Components of the Model

16.2.1 Prior Probability: The Class Frequency

We begin with the simplest quantity in the Naive Bayes model: the prior probability. The prior $p_Y(C_k)$ represents how frequently a class appears in the dataset before we look at any feature values. Since the class label Y is a discrete random variable, we use a probability mass function to estimate this value from the training data.

If we let count(C_k) denote the number of samples whose label is C_k, and let count(Ω) be the total number of samples in the dataset, then the empirical prior can be estimated as

$$p_Y\left(C_k\right)=\frac{\operatorname{count}\left(C_k\right)}{\operatorname{count}\left(\Omega\right)},\quad k=1,2,3 \tag{1}$$

This ratio simply measures what fraction of the training samples belong to class C_k.

For example, in the Iris dataset, there are 150 samples in total—50 setosa, 50 versicolor, and 50 virginica. Because each class has the same number of samples, the estimated prior for all three classes is

$$p_Y\left(C_k\right)=\frac{50}{150}=\frac{1}{3},\quad k=1,2,3 \tag{2}$$

This tells us that, before considering any features, each class is equally likely.

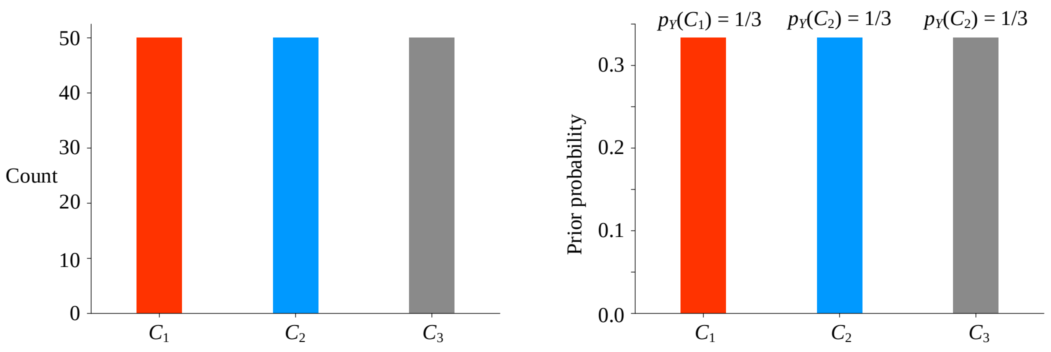


Figure 2. Class frequencies and estimated prior probabilities in the Iris dataset

16.2.2 Likelihood: How Features Fit Each Class

Earlier, we mentioned that the word “naive” in Naive Bayes comes from a key assumption: the features are conditionally independent given the class label. This assumption greatly simplifies how we estimate the likelihood, which refers to the class-conditional probability density $f_{X1,X2|Y}\left(x_1,x_2|C_k\right)$.

Under the conditional independence assumption, the joint density of the features factorizes into a product of one-dimensional densities:

$$\underbrace{f_{X1,X2|Y}\left(x_1,x_2|C_k\right)}_{\text{Likelihood}}=\underbrace{f_{X1|Y}\left(x_1|C_k\right)f_{X2|Y}\left(x_2|C_k\right)}_{\text{Conditional independence}} \tag{3}$$

This is the core simplification that makes Naive Bayes computationally efficient.

To estimate each one-dimensional likelihood term, we often assume that continuous features follow a Gaussian (normal) distribution under each class. With this choice, the classifier is more precisely called a Gaussian Naive Bayes classifier.

When the conditional distributions $f_{X1\,|\,Y}(x_1\,|\,C_k)$ and $f_{X2\,|\,Y}(x_2\,|\,C_k)$ are modeled by a Gaussian (normal) distribution, the classifier is called a Gaussian Naive Bayes classifier.

Each feature in each class is represented by a one-dimensional Gaussian curve defined by its own mean and variance, which are estimated from the training data. Figure 3 illustrates this idea: given the class label $Y = C_1$, we model the marginal conditional probabilities $f_{X1\,|\,Y}(x_1\,|\,C_1)$ and $f_{X2\,|\,Y}(x_2\,|\,C_1)$ as Gaussian curves.

By multiplying these two one-dimensional likelihoods, we obtain the two-dimensional likelihood surface $f_{X1,X2\,|\,Y}(x_1, x_2\,|\,C_1)$. Because of the independence assumption, the resulting contour lines form axis-aligned ellipses.

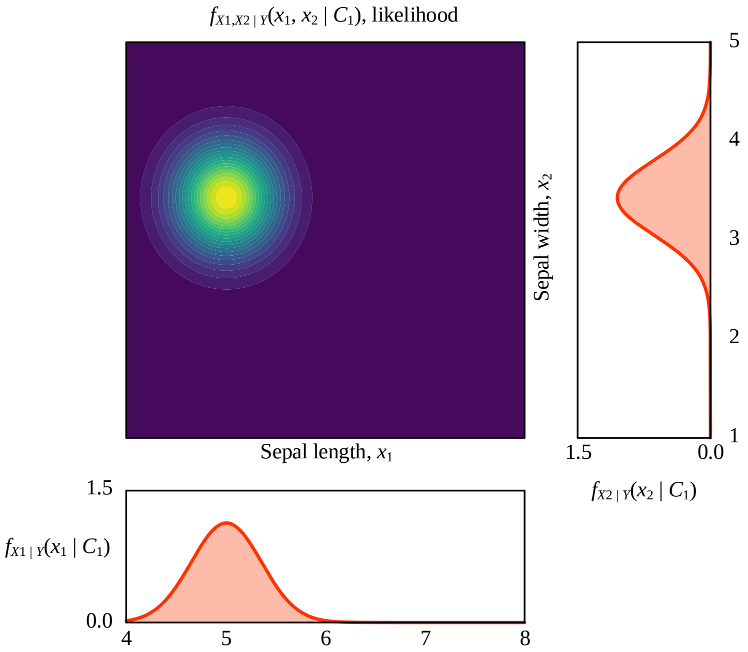


Figure 3. Estimating the likelihood $f_{X1,X2\,|\,Y}(x_1, x_2\,|\,C_1)$ under the conditional independence assumption

To visualize this more concretely, Figure 4 shows the corresponding likelihood surface and contour plot for class C_1. The surface represents the value of $f_{X1,X2 \mid Y}(x_1, x_2 \mid C_1)$ over the feature space, and the volume under the surface equals 1, as expected for a probability density function.

In Figure 4 (b), the red points represent real training samples belonging to class C_1. The smooth Gaussian surface provides an idealized description of how those samples are distributed in the space of (x_1, x_2).

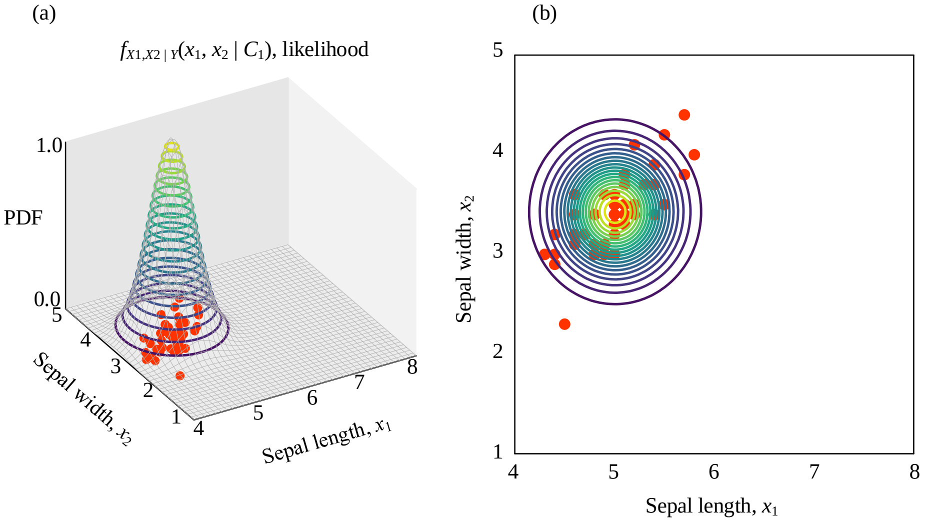


Figure 4. Likelihood surface and contour of $f_{X1,X2 \mid Y}(x_1, x_2 \mid C_1)$ assuming conditional independence

The same process can be repeated for the other classes. Figure 5 and Figure 6 show the estimated likelihoods for class C_2. Each feature, X_1 and X_2, has its own Gaussian distribution conditioned on $Y = C_2$, and their product gives the joint likelihood $f_{X1,X2 \mid Y}(x_1, x_2 \mid C_2)$.

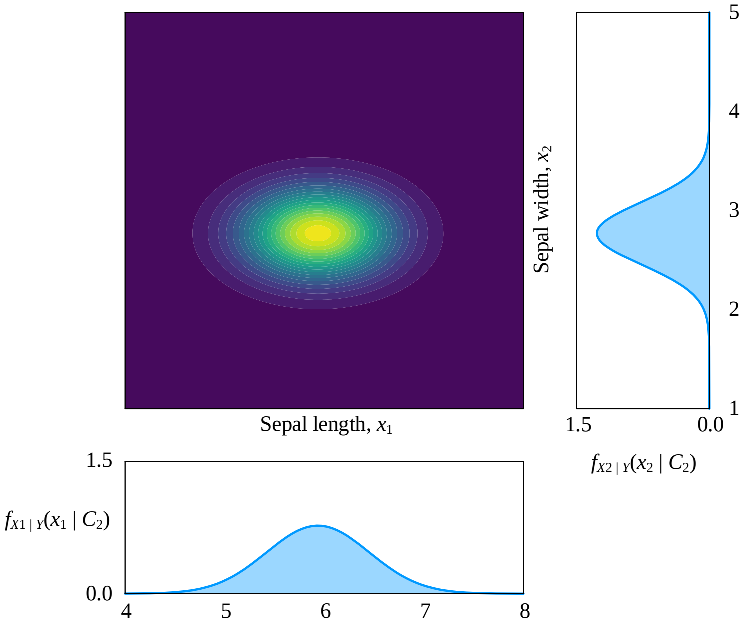


Figure 5. Estimating the likelihood $f_{X1,X2 \mid Y}(x_1, x_2 \mid C_2)$ under the conditional independence assumption

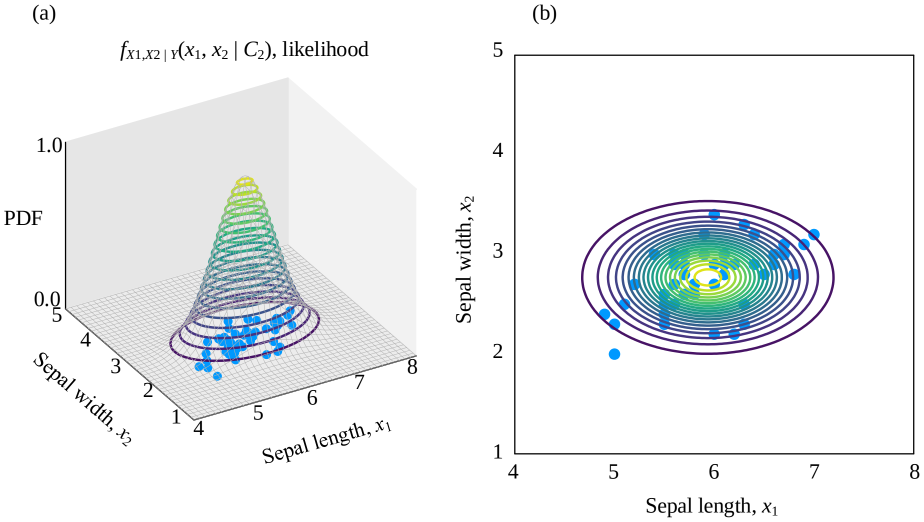


Figure 6. Likelihood surface and contour of $f_{X1,X2 \mid Y}(x_1, x_2 \mid C_2)$ assuming conditional independence

Likewise, Figure 7 and Figure 8 illustrate the estimated likelihoods for class C_3. By examining these Figures side by side, we can see how each class produces its own distinct likelihood surface—each shaped by the distribution of its data points. Together, these likelihood functions form the foundation for computing the posterior probabilities used in classification.

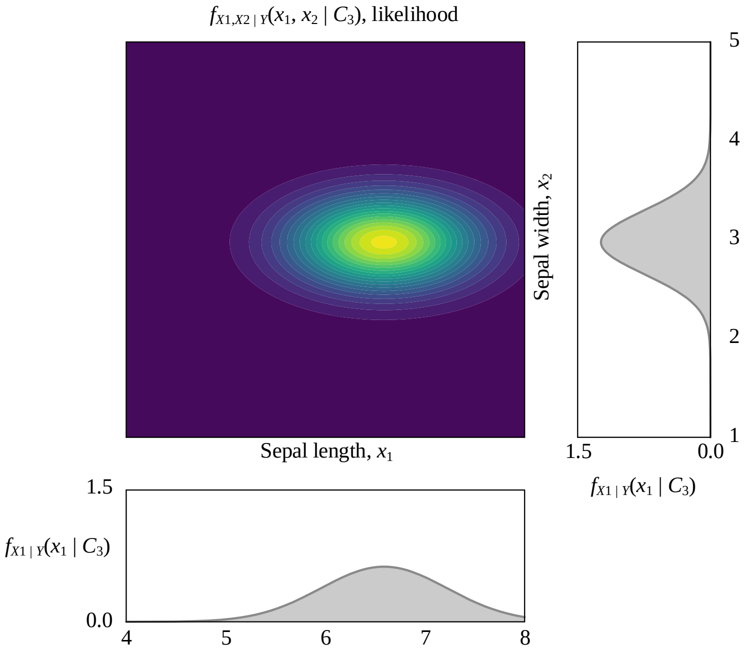


Figure 7. Estimating the likelihood $f_{X1,X2 \mid Y}(x_1, x_2 \mid C_3)$ under the conditional independence assumption

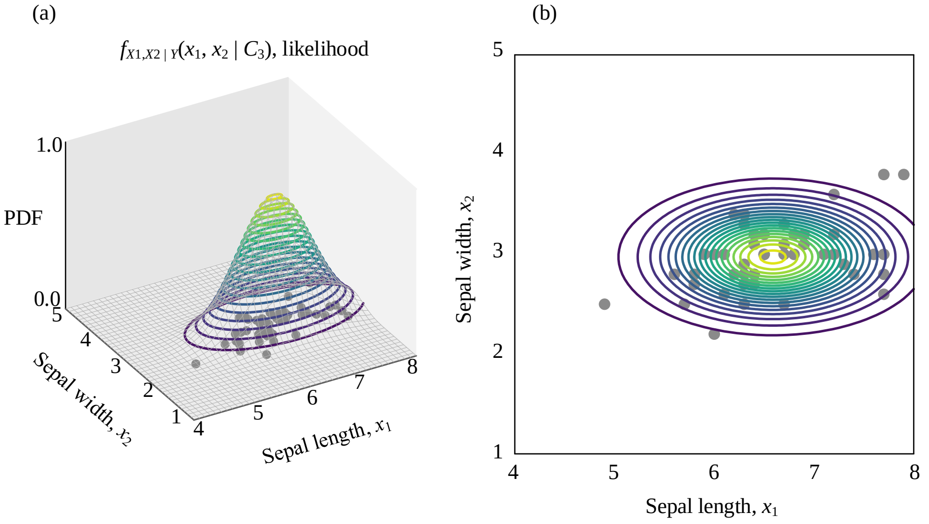


Figure 8. Likelihood surface and contour of $f_{X1,X2 \mid Y}(x_1, x_2 \mid C_3)$ assuming conditional independence

16.2.3 Evidence: The Normalizing Factor

Once we have the class priors and the class-conditional likelihoods, the next quantity to compute is the evidence (also called the marginal likelihood) for the observed features. The evidence is the probability density of observing that feature vector regardless of class and is given by summing the class-conditional densities weighted by their priors:

$$\begin{aligned} \underbrace{f_{X1,X2}(x_1,x_2)}_{\text{Evidence}} &= \underbrace{f_{X1,X2,Y}(x_1,x_2,C_1)}_{\text{Joint}} + \underbrace{f_{X1,X2,Y}(x_1,x_2,C_2)}_{\text{Joint}} + \underbrace{f_{X1,X2,Y}(x_1,x_2,C_3)}_{\text{Joint}} \\ &= \underbrace{p_Y(C_1)}_{\text{Prior}} \underbrace{f_{X1,X2|Y}(x_1,x_2|C_1)}_{\text{Likelihood}} + \underbrace{p_Y(C_2)}_{\text{Prior}} \underbrace{f_{X1,X2|Y}(x_1,x_2|C_2)}_{\text{Likelihood}} + \underbrace{p_Y(C_3)}_{\text{Prior}} \underbrace{f_{X1,X2|Y}(x_1,x_2|C_3)}_{\text{Likelihood}} \end{aligned} \tag{4}$$

This formula shows that the evidence is a weighted average of the class likelihoods, where each weight is the prior probability of the corresponding class. Intuitively, the evidence answers the question “how likely is it to see a feature vector like (x_1, x_2) in the whole dataset?”

It serves two roles: mathematically it is the normalizing constant in Bayes’ theorem, and conceptually it is a standalone estimate of the feature density across all classes. Because it is a probability density, the surface defined by $f_{X1,X2}(x_1, x_2)$ integrates to one over the feature space.

Figure 9 displays contour lines of the evidence together with the marginal distributions of each feature; this offers a compact view of where the dataset places most of its probability mass.

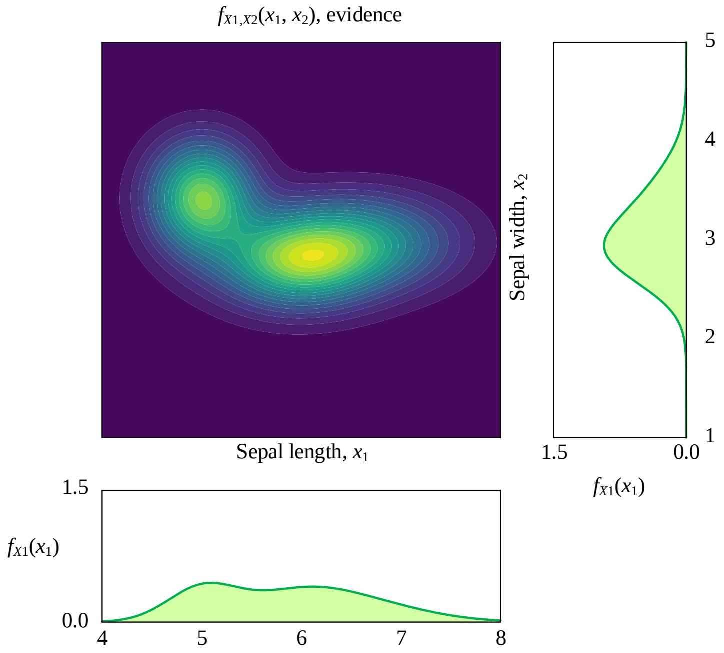


Figure 9. Contour plot of the evidence $f_{X1,X2}(x_1, x_2)$ with marginal feature distributions.

Figure 10 shows the same evidence as a three-dimensional surface and contour plot, emphasizing that the evidence surface is formed by blending the class-conditional surfaces according to their priors.

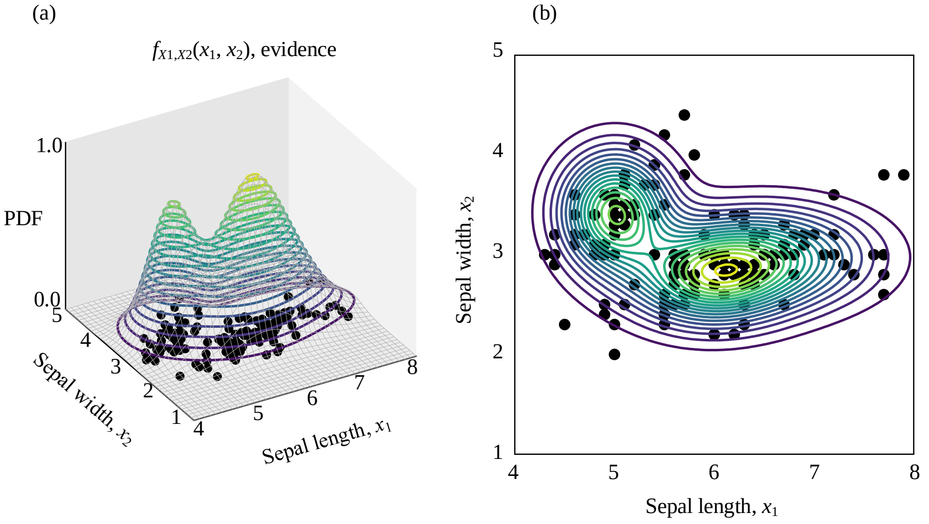


Figure 10. Evidence surface and contour plot for $f_{X1,X2}(x_1, x_2)$: the marginal feature density across all classes.

Figure 11 is a schematic illustration of the computation: each class-conditional likelihood contributes a surface, each surface is scaled by its class prior, and the sum of those scaled surfaces produces the overall evidence surface.

Understanding the evidence is useful beyond normalization. Comparing evidence values at different feature points helps detect outliers: unusually low evidence suggests the feature vector is unlikely under the model. The evidence is also central to model comparison and marginal likelihood-based techniques because it quantifies how well the model explains the observed features.

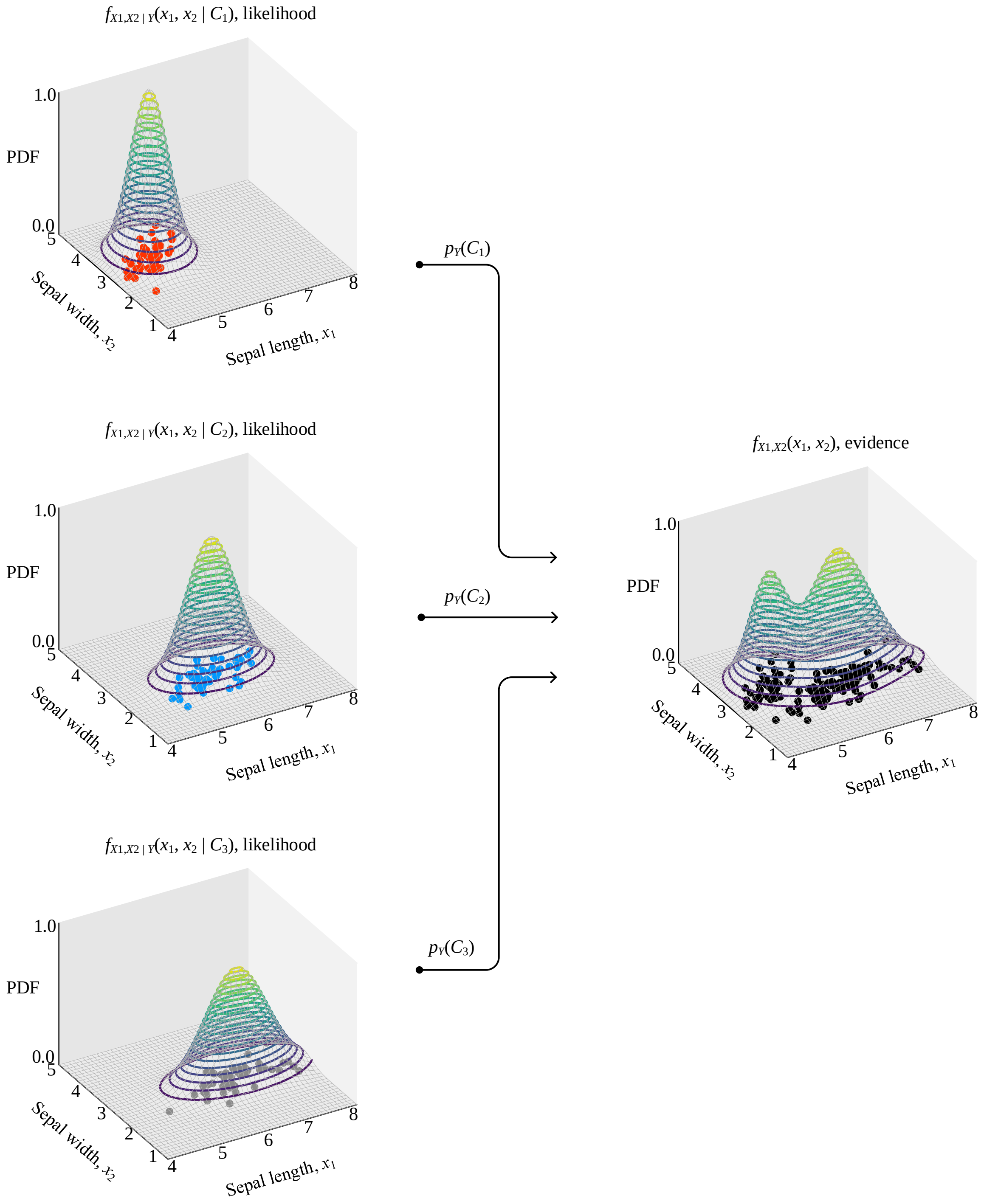


Figure 11. Priors weight each class-conditional likelihood to form the evidence surface $f_{X1,X2}(x_1, x_2)$

In the next section we will combine the evidence with priors and likelihoods to compute the posterior probabilities used for classification, and we will show how the evidence cancels out when comparing classes, which simplifies decision rules in many practical implementations.

16.3 Computing the Posterior Probability

16.3.1 Posterior Formula and Interpretation

Once we have the prior and likelihood, we can compute the posterior probability, which is the final quantity used to make a prediction in Bayesian classification. The posterior probability $f_{Y|X1,X2}(C_1 \mid x_1, x_2)$ answers a simple and intuitive question: Given the observed features (x_1, x_2), how likely is it that the sample belongs to class C_1?

Unlike likelihood functions or density estimates, the posterior is a true probability, not a probability density. Therefore, its value always lies between 0 and 1, inclusive. By Bayes' theorem, when the evidence term $f_{X1,X2}(x_1, x_2)$ is non-zero, the posterior can be computed as:

$$\underbrace{f_{Y|X1,X2}\left(C_1 | x_1, x_2\right)}_{\text{Posterior}} = \frac{\overbrace{f_{X1,X2,Y}\left(x_1, x_2, C_1\right)}^{\text{Joint}}}{\underbrace{f_{X1,X2}\left(x_1, x_2\right)}_{\text{Evidence}}} \tag{5}$$

Figure 12 shows the resulting posterior distribution. Because it represents a probability, the surface ranges only from 0 to 1. Sometimes, this value is also called a **membership value**, since it reflects the degree to which the sample belongs to a particular class.

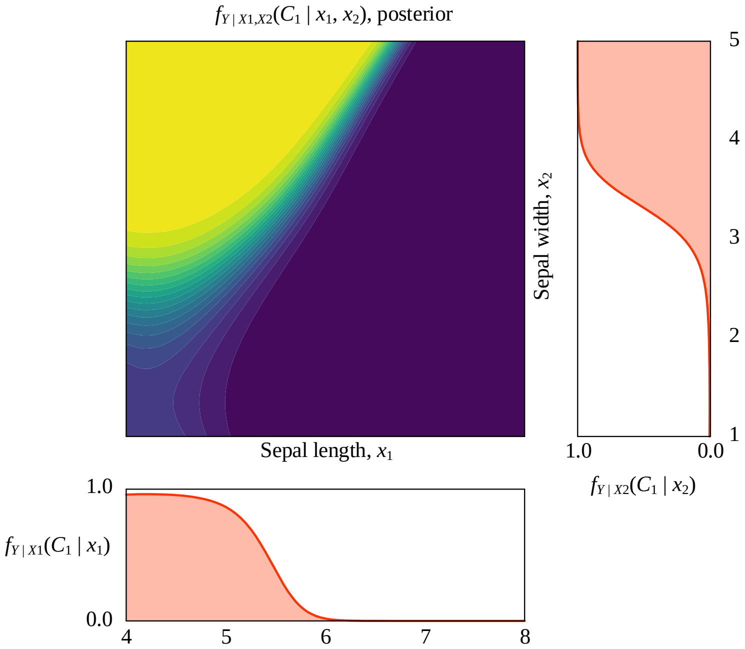


Figure 12. Contour plot of the posterior $f_{Y|X1,X2}(C_1 \mid x_1, x_2)$ with marginal feature distributions.

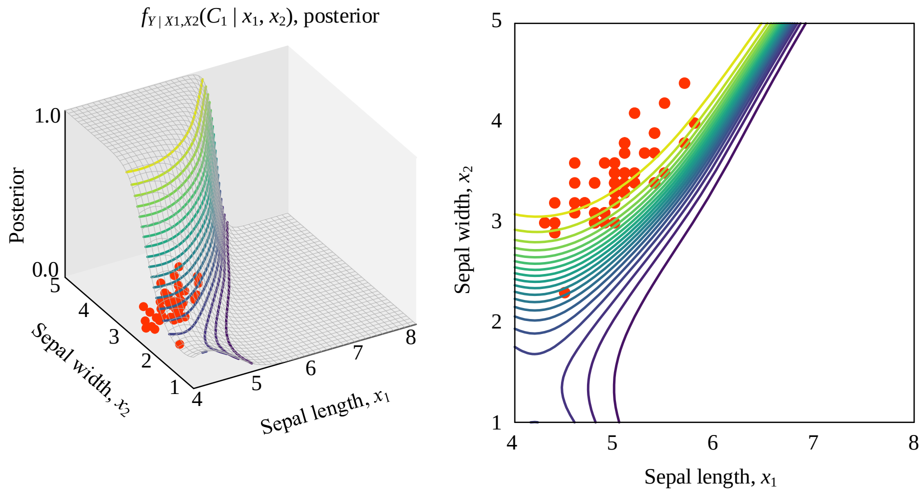


Figure 13. Posterior surface and contour plot for $f_{Y|X1,X2}(C_1 | x_1, x_2)$

Using the same approach, we can compute the posterior probabilities for C_2 and C_3:

$$\underbrace{f_{Y|X1,X2}\left(C_2|x_1,x_2\right)}_{\text{Posterior}} = \frac{\overbrace{f_{X1,X2,Y}\left(x_1,x_2,C_2\right)}^{\text{Joint}}}{\underbrace{f_{X1,X2}\left(x_1,x_2\right)}_{\text{Evidence}}}$$
$$\underbrace{f_{Y|X1,X2}\left(C_3|x_1,x_2\right)}_{\text{Posterior}} = \frac{\overbrace{f_{X1,X2,Y}\left(x_1,x_2,C_3\right)}^{\text{Joint}}}{\underbrace{f_{X1,X2}\left(x_1,x_2\right)}_{\text{Evidence}}} \tag{6}$$

Figure 14 ~ Figure 17 visualize these posterior surfaces and their contour plots under the naïve Bayes assumption (conditional independence of features). Because we are dealing with a three-class problem, the posterior probabilities at any point must satisfy:

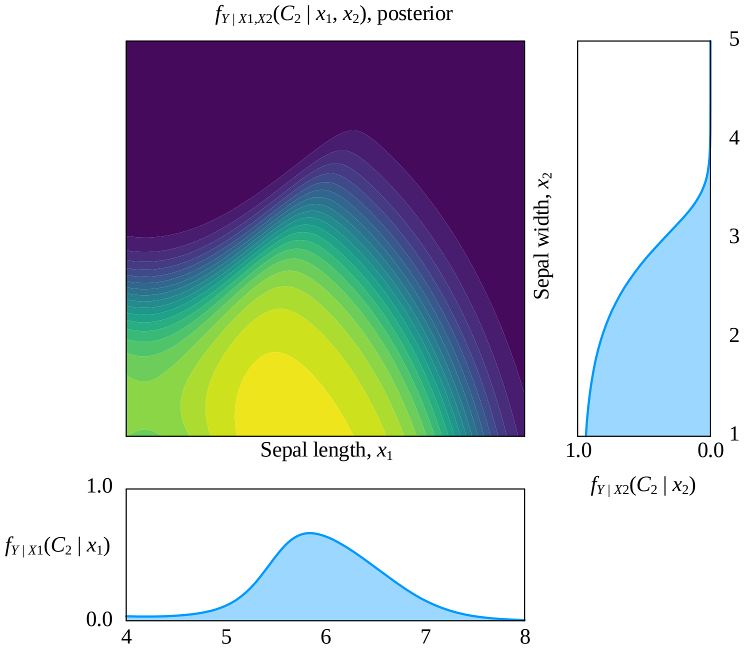


Figure 14. Contour plot of the posterior $f_{Y|X1,X2}(C_2 | x_1, x_2)$ with marginal feature distributions.

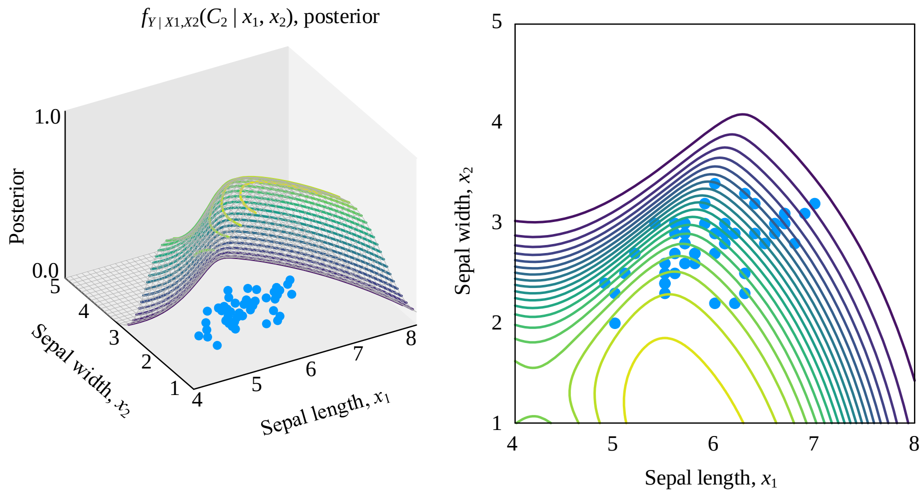


Figure 15. Posterior surface and contour plot for $f_{Y|X1,X2}(C_2 | x_1, x_2)$

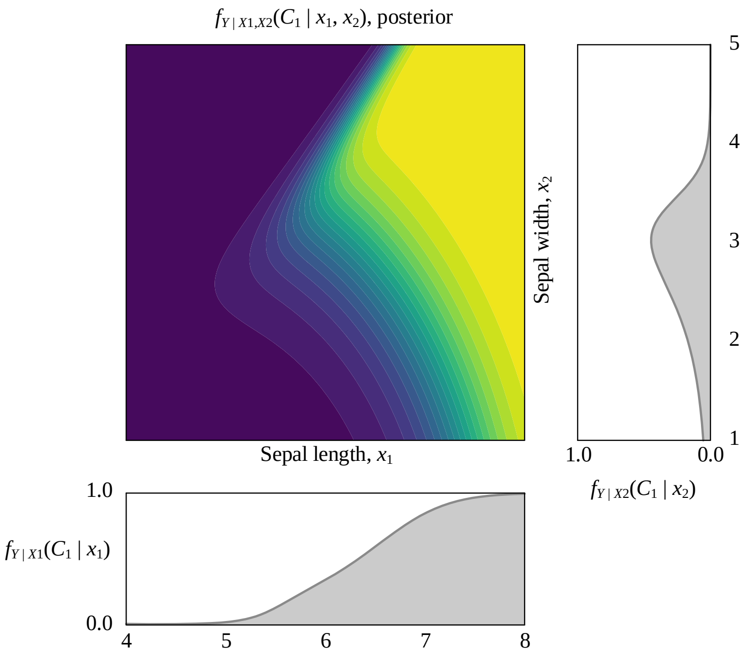


Figure 16. Contour plot of the posterior $f_{Y|X1,X2}(C_3 | x_1, x_2)$ with marginal feature distributions.

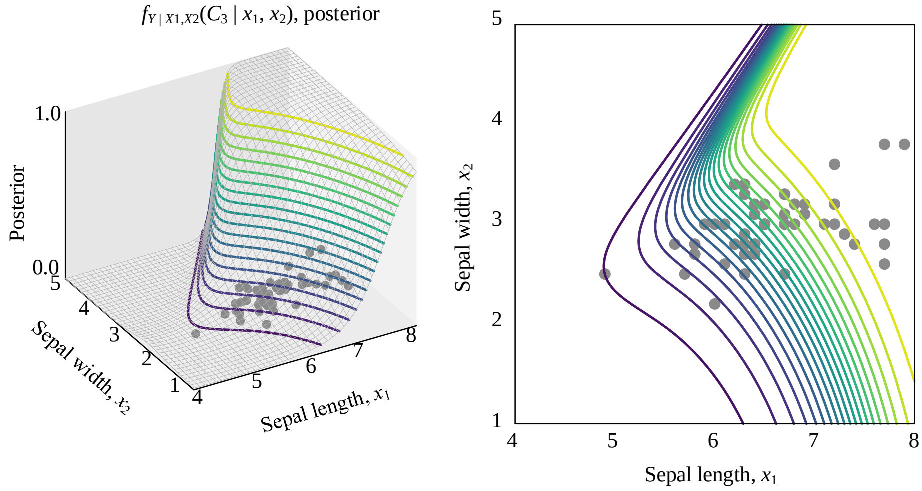


Figure 17. Posterior surface and contour plot for $f_{Y \mid X1,X2}(C_3 \mid x_1, x_2)$

16.3.3 Simplifying the Decision Rule

Because we are dealing with a three-class problem, the posterior probabilities at any point must satisfy:

$$\underbrace{f_{Y|X1,X2}\left(C_1|x_1,x_2\right)}_{\text{Posterior}}+\underbrace{f_{Y|X1,X2}\left(C_2|x_1,x_2\right)}_{\text{Posterior}}+\underbrace{f_{Y|X1,X2}\left(C_3|x_1,x_2\right)}_{\text{Posterior}}=1 \tag{7}$$

This means that at every point in the feature plane, the sample must belong to one—and only one—of the three classes. Geometrically, if you stack the three posterior surfaces (Figure 13 (a), Figure 15 (a), and Figure 17 (a)), their heights always sum to exactly 1.

One more important observation comes from comparing the Bayes formulas: the denominator (the evidence) is identical for all three classes. Therefore, to predict the class label, we do **not** actually need to compute the posterior probability explicitly. It is mathematically equivalent to choosing the class with the largest product of likelihood and prior:

$$\begin{cases}\underbrace{f_{X1,X2,Y}\left(x_1,x_2,C_1\right)}_{\text{Joint}}=\underbrace{p_Y\left(C_1\right)}_{\text{Prior}}\underbrace{f_{X1,X2|Y}\left(x_1,x_2|C_1\right)}_{\text{Likelihood}}\\ \underbrace{f_{X1,X2,Y}\left(x_1,x_2,C_2\right)}_{\text{Joint}}=\underbrace{p_Y\left(C_2\right)}_{\text{Prior}}\underbrace{f_{X1,X2|Y}\left(x_1,x_2|C_2\right)}_{\text{Likelihood}}\\ \underbrace{f_{X1,X2,Y}\left(x_1,x_2,C_3\right)}_{\text{Joint}}=\underbrace{p_Y\left(C_3\right)}_{\text{Prior}}\underbrace{f_{X1,X2|Y}\left(x_1,x_2|C_3\right)}_{\text{Likelihood}}\end{cases} \tag{8}$$

This leads directly to the decision boundary seen in the Iris classification example (Figure 18), based on a Gaussian naïve Bayes model.

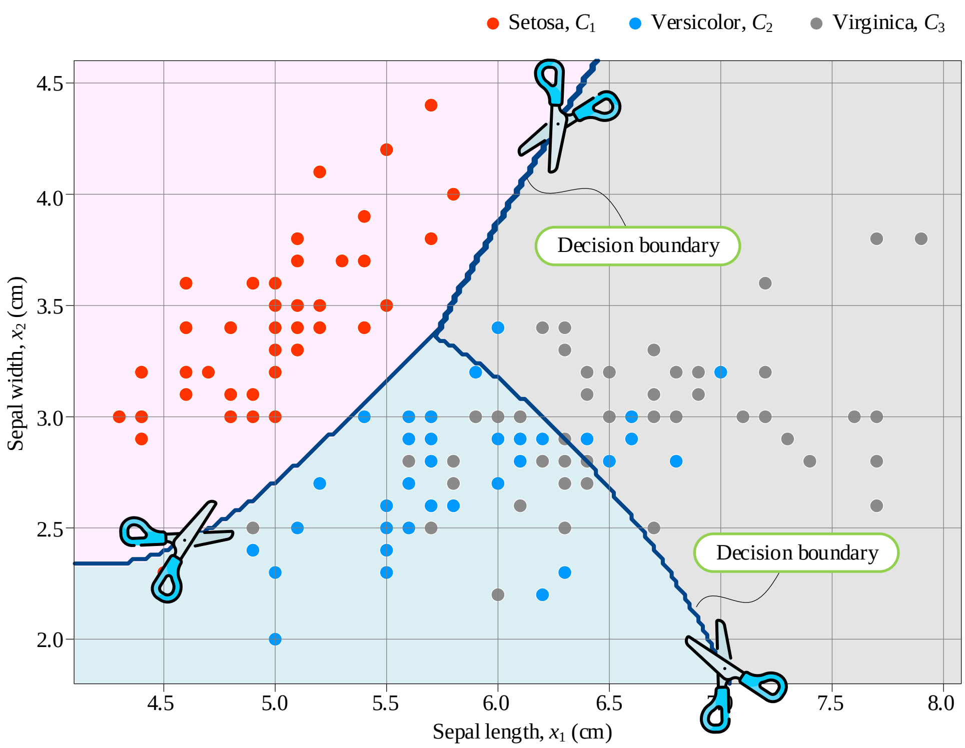


Figure 18. Decision boundary of Gaussian naïve Bayes for the Iris classification task. Figure generated by Ch16_01_Gaussian_Naive_Bayes.ipynb.

16.4 Conclusion

The Naive Bayes classifier is a simple probabilistic model that predicts class labels using Bayes theorem. It estimates how common each class is in the data and how likely the features are for each class. The method is called naive because it assumes that all features are independent once the class label is known. This assumption makes the model fast and easy to train, even when there are many features. After learning the prior and the likelihood from the training set, the classifier computes the posterior probability for each class and chooses the class with the highest value.

Although it is based on a strong assumption, Naive Bayes often performs surprisingly well in practice and is widely used as a baseline, especially for text and high dimensional data. Variants such as Gaussian, Multinomial, and Bernoulli Naive Bayes handle different types of feature distributions.