Bayesian Classification*

February 23, 2022

1 Probability Basics

1.1 Introduction

A key concept in the field of pattern recognition is that of uncertainty. It arises both through the noise on measurements, as well as through the finite size of data sets. Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations for pattern recognition. It allows us to make optimal predictions given all the information available to us, even though that information may be incomplete or ambiguous.

1.2 An Example

We will introduce the basic concepts of probability theory by considering a simple example. Imagine we have two boxes, one red and one blue, and in the red box we have 2 apples and 6 oranges, and in the blue box, we have 3 apples and 1 orange (Figure 1). Now suppose we randomly pick one of the boxes, and from that box, we randomly select an item of fruit. Having observed which sort of fruit it is, we put it back in the box from which it came. The next time we draw another fruit from one of the boxes, the probability distribution remains the same as this time. This is called i.i.d (i.e., identical independent distribution) in probability theory. We could repeat this process many times. Let us suppose that in doing so we pick the red box 40% of the time and we pick the blue box 60% of the time.

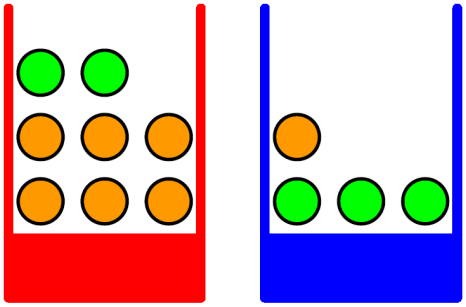


Figure 1: Boxes with fruits (green color: apple; orange color: orange).

*References

- Christopher Bishop. Pattern Recognition and Machine Learning. 2006
- Sergios Theodoridis, Konstantinos Koutroumbas. Pattern Recognition. 2009

In this example, the identity of the box that will be chosen is a random variable, which we denote by B. This random variable can take two possible values, r (corresponding to the red box) or b (corresponding to the blue box). The identity of the fruit is also a random variable and will be denoted by F. It can take either of the values a (for apple) or o (for orange).
We define the probability of an event to be the fraction of times that event occurs out of the total number of trials, in the limit that the total number of trials goes to infinity. Thus the probability of selecting the red box is 4/10, and the probability of selecting the blue box is 6/10, i.e.,

$$P(B = r) = \frac{4}{10},\ P(B = b) = \frac{6}{10}.$$

By definition, probabilities must lie in the interval [0, 1]. Also, if the events are mutually exclusive and if they include all possible outcomes (for instance, in this example the box must be either red or blue), then we see that the probabilities for those events must sum to one.

1.3 Probability Rules

Let's now ask questions such as: "what is the overall probability that we will pick an apple?", or "given that we have chosen an orange, what is the probability that the box we chose was the blue one?". We can answer such questions, and indeed much more complex questions associated with problems in pattern recognition, once we have equipped ourselves with the two fundamental rules of probability, known as the *sum rule* and the *product rule*.

$$P(X) = \sum_{Y} P(X, Y)$$

$$P(X, Y) = P(Y|X) \cdot P(X)$$

$P(X, Y)$ is the joint probability that X and Y happen at the same time. $P(X)$ is simply the probability of X, which can be obtained by marginalizing $P(X, Y)$ over all possible Ys. $P(Y|X)$ is a conditional probability: given that X happens, the probability that Y will happen. These two simple rules form the basis for all of the probabilistic machinery that we use throughout this course.
Returning back to the example of boxes of fruits, we can obtain

$$P(F = a|B = r) = \frac{1}{4}$$

$$P(F = o|B = r) = \frac{3}{4}$$

$$P(F = a|B = b) = \frac{3}{4}$$

$$P(F = o|B = b) = \frac{1}{4}$$

Now we can use the sum rule and product rule to evaluate the overall probability of choosing an apple

$$\begin{aligned} P(F = a) &= P(F = a|B = r) \cdot P(B = r) + P(F = a|B = b) \cdot P(B = b) \\ &= \frac{1}{4} \cdot \frac{4}{10} + \frac{3}{4} \cdot \frac{6}{10} \\ &= \frac{11}{20} \end{aligned}$$

Similarly, we can obtain $P(F = o) = 9/20$. It is easily observed that $P(F = a) + P(F = o) = 1$, which means that when we pick a random fruit from a random box, we either end up an apple or orange, no other scenarios.
Finally, if the joint distribution of two variables factorizes into the product of the marginals

$$P(X, Y) = P(X) \cdot P(Y),$$

X and Y are said to be independent. From the product rule, we see that $P(Y|X) = P(Y)$, and so the conditional distribution of Y given X is indeed independent of the value of X. In our boxes of fruit example, if each box contained the same fraction of apples and oranges, then $P(F|B) = P(F)$, so that the probability of selecting an apple/orange is independent of which box is chosen.

1.4 Probability Density Function

As well as considering probabilities defined over discrete sets of events, we also wish to consider probabilities with respect to continuous variables. We shall limit ourselves to a relatively informal discussion. If the probability of a real-valued variable X falling in the interval $[X, X + \delta X]$ is given by $p(X) \cdot \delta X$ for $\delta X \to 0$, then $p(X)$ is called the probability density over X. This is illustrated in Figure 2. The probability that X will lie in an interval $[a, b]$ is then given by

$$p(X \in (a, b)) = \int_a^b p(X)\, dX.$$

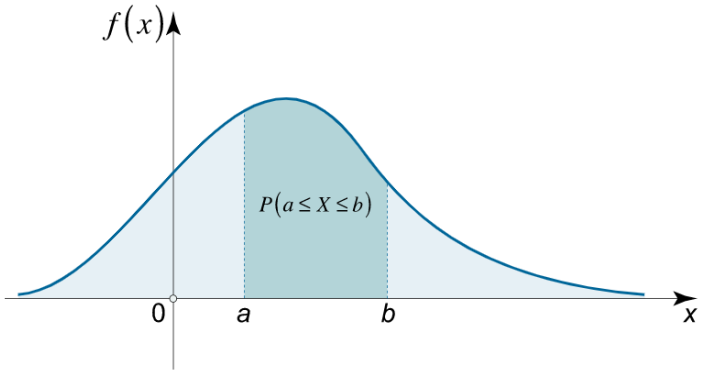


Figure 2: Probability density function.[1]

2 Bayesian Classification

From the product rule, together with the symmetry property $P(X, Y) = P(Y, X)$, we immediately obtain the following relationship between conditional probabilities[2]

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}, \tag{1}$$

which is called **Bayes' theorem** and which plays a central role in pattern recognition and machine learning.

[1] https://math24.net/probability-density-function.html
[2] Bayes' theorem. Wikipedia

2.1 Standard Bayesian Classification

We will focus on the two-class case. Let y_1, y_2 be the two classes to which our patterns belong. In the sequel, we assume that the prior probabilities $P(y_1)$, $P(y_2)$ are known. This is a very reasonable assumption because even if they are not known, they can easily be estimated from the available training samples. If N is the total number of available training patterns, and N_1, N_2 of them belong to y_1 and y_2, respectively, then

$$P(y_1) \approx N_1/N$$
$$P(y_2) \approx N_2/N$$

The other statistical quantities assumed to be known are the class conditional probability density distribution $p(\mathbf{x}|y_i)$, where $i = 1, 2$, describing the distribution of the feature vectors in each of the classes. If these are not known, they can also be estimated from the available training data. In the case that feature vectors can take only discrete values, density functions $p(\mathbf{x}|y_i)$ become probabilities and will be denoted by $P(\mathbf{x}|y_i)$.
Applying the Bayes' theorem, we have

$$P(y_i|\mathbf{x}) = \frac{p(\mathbf{x}|y_i)P(y_i)}{p(\mathbf{x})},$$

where $p(\mathbf{x})$ is the input data probability distribution and for which we have

$$p(\mathbf{x}) = \sum_{i=1}^{2} p(\mathbf{x}|y_i)P(y_i).$$

The Bayes classification rule can now be stated as:

- if $P(y_1|\mathbf{x}) < P(y_2|\mathbf{x})$, assign sample $\mathbf{x}$ to y_2;
- if $P(y_1|\mathbf{x}) > P(y_2|\mathbf{x})$, assign sample $\mathbf{x}$ to y_1;
- if the two quantities are equal, the pattern can be assigned to any of the classes.

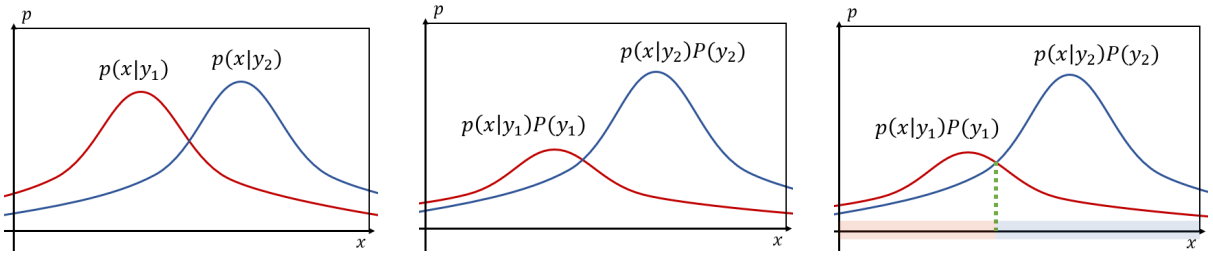
Applying the Bayes rule and eliminating $p(\mathbf{x})$ for all classes, we can determine the class of a sample by considering the inequality between

$$p(\mathbf{x}|y_1)P(y_1) \text{ and } p(\mathbf{x}|y_2)P(y_2).$$

If $P(y_1) = P(y_2) = 1/2$, then we determine the class of a sample by considering the inequality between

$$p(\mathbf{x}|y_1) \text{ and } p(\mathbf{x}|y_2).$$

Figure 3 summarizes the steps we take to perform Bayes classification.



(a) Step 1 (b) Step 2 (c) Step 3

Figure 3: Steps to perform Bayes classification. (a) Step 1: compute class conditional probabilities. (b) Step 2: multiply with class priors. (c) Step 3: obtain the class posterior probabilities and find the decision boundary.

2.2 Minimizing the Average Risk

Standard Bayes classification is not always the best criterion to be adopted for minimization. This is because it assigns the same importance to all errors. However, there are cases in which some wrong decisions may have more serious implications than others. For example, it is much more serious for a doctor to make a wrong decision and a malignant tumor to be diagnosed as a benign one, than the other way around. If a benign tumor is diagnosed as a malignant one, the wrong decision will be cleared out during subsequent clinical examinations. However, the results from the wrong decision concerning a malignant tumor may be fatal. In such cases, it is more appropriate to assign a penalty term to weigh each error.
Considering the two-case classification problem, we introduce the loss matrix

$$l = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix},$$

where λ_{ij} denotes the loss of assigning a sample of class i to class j. Apparently, we have $\lambda_{11} = \lambda_{22} = 0$.
We compute the risk of assigning a sample $\mathbf{x}$ to the two classes

$$\begin{aligned} l_1 &= \lambda_{11} \cdot p(\mathbf{x}|y_1) \cdot P(y_1) + \lambda_{21} \cdot p(\mathbf{x}|y_2) \cdot P(y_2) \\ l_2 &= \lambda_{12} \cdot p(\mathbf{x}|y_1) \cdot P(y_1) + \lambda_{22} \cdot p(\mathbf{x}|y_2) \cdot P(y_2) \end{aligned}.$$

We assign $\mathbf{x}$ to class y_1 if $l_1 < l_2$, which means

$$\lambda_{21} \cdot p(\mathbf{x}|y_2) \cdot P(y_2) < \lambda_{12} \cdot p(\mathbf{x}|y_1) \cdot P(y_1).$$

2.3 Bayes Error

Recall the decision boundary presented in Figure 3 (c), the dotted line is a threshold partitioning the feature space into two regions, R_1 and R_2. According to the Bayes decision rule, for all values of $\mathbf{x}$ in R_1 the classifier decides y_1 and for all values in R_2 it decides y_2. However, it is obvious from the figure that decision errors are unavoidable. Indeed, there is a finite probability for an $\mathbf{x}$ to lie in the R_2 region and at the same time to belong in class y_1. Then our decision is in error. The total probability, P_e, of committing a decision error for the case of two classes, is given by

$$P_e = \int_{-\infty}^{x_0} p(x|y_2)P(y_2)\,dx + \int_{x_0}^{\infty} p(x|y_1)P(y_1)\,dx.$$

Here we use x instead of the bold $\mathbf{x}$, indicating that the input sample is a scalar value instead of a multi-dimensional vector. P_e is equal to the total shaded area under the curves in Figure 4.

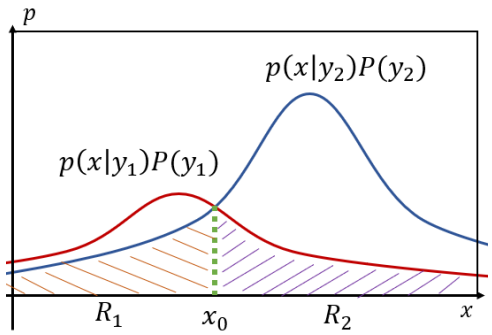


Figure 4: Bayes error.

Bayes error is a very important concept in machine learning. It is the minimum error you could obtain using any kind of classifiers (e.g., SVM, RF, Deep neural networks). You could imagine obtaining another decision boundary to divide the classes, and no other boundary will achieve a smaller error (i.e., the shaded areas) than the dotted line we denote in Figure 4. However, Bayes error is a theoretical value that cannot be obtained from real-world problems, as we don't have the exact data distributions of various classes in the real world.