Histograms and Bayes Classifiers

Raúl Rojas
Freie Universität Berlin

July 2014

A histogram is a representation of the distribution of data points according to one or more variables. Assume that we have a pond with N_A fishes of class A and N_B fishes of class B. If we take one fish out of the pond and record its class and length (the selected feature), we could build the graphical representation shown in Fig. 1. Here, the lengths of the fishes are collected, for example, in "bins" 10 cm wide.

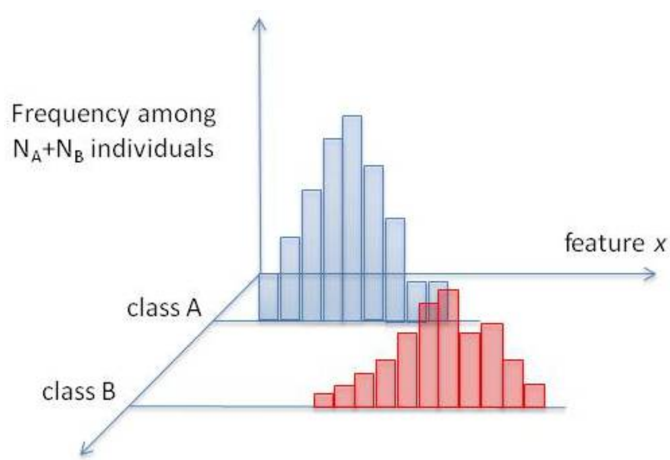

Figure 1: The histograms of two classes of fish arranged according to length.

We call $P(A) = N_A/(N_A + N_B)$ the relative frequency of fishes of class A in the pond, while $P(B) = N_B/(N_A + N_B)$ represents the relative frequency of class B. These are the *a priori* probabilities of finding a fish of class A or class B when pulling a fish randomly out of the pond. If there are, for example, 60% of fishes of class A and 40% of class B, then $P(A) = 0.6$ and $P(B) = 0.4$.

If we only consider all the recorded fishes of class A , we can compute the *density* of the probability distribution of the class A along the variable x. We call this $P(x|A)$ because given that we are talking about class A, we can see in the graph of $P(x|A)$ how often the bin corresponding to x is present in the data. We can do the same for class B.

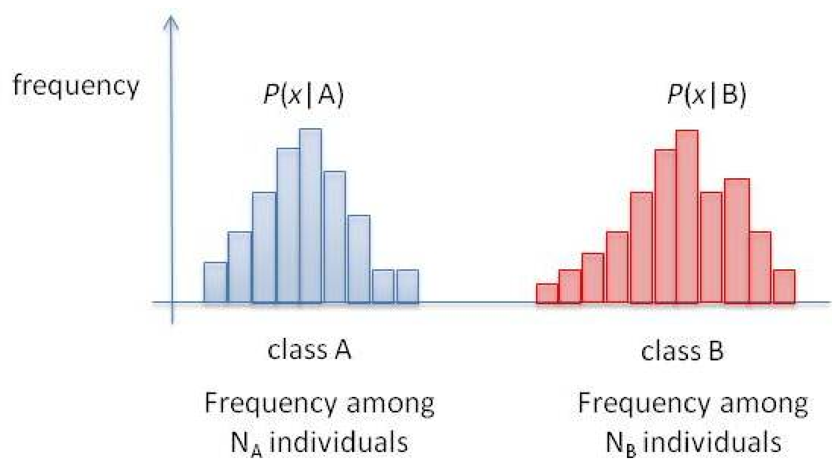

Figure 2: The histograms of two classes considered independently of each other. The frequencies of the data add to 100% for each histogram.

As we see in Fig. 2 each partial histogram covers only class A, resp. class B. The surface under the histograms has value one (100% of all individuals of class A, or of class B).

Using this convention we immediately see that the plot in Fig. 1 corresponds to the shape of $P(x|A)P(A)$ and of $P(x|B)P(B)$. Since $P(x \cap A) = P(x|A)P(A)$ and $P(x \cap B) = P(x|B)P(B)$, what we see in Fig. 1 is the common histogram of all fishes distributed into two classes.

In pattern recognition we are interested in finding the class of a given measurement x (that is, we want to assign a fish to a class based only on its length). In our example, we are interested in comparing $P(A|x)$ and $P(B|x)$. We select A if $P(A|x) \geq P(B|x)$, and B otherwise.

According to probability theory

$$\begin{aligned} P(A|x)P(x) &= P(A \cap x) \\ \text{or equivalently} \quad P(A|x) &= P(x|A)(P(A)/P(x)) \end{aligned}$$

and similarly $P(B|x) = P(x|B)(P(B)/P(x))$. This is called the Bayes formula. It allows us to infer classifications by working with probabilities "in reverse". We select class A whenever

$$\begin{aligned} P(A|x) &\geq P(B|x)) \\ \frac{P(x|A)P(A)}{P(x)} &\geq \frac{P(x|B)P(B)}{P(x)} \\ \Rightarrow \quad P(x|A)P(A) &\geq P(x|B)P(B) \end{aligned} \tag{1}$$

As we see, it is not necessary to know the exact value of $P(x)$ for a comparison. The two sides in the expression above are the readings from the histograms in Fig. 1. Therefore, Fig. 3 is just Fig. 1 redrawn, but now in such a way as to use the notation just introduced above.

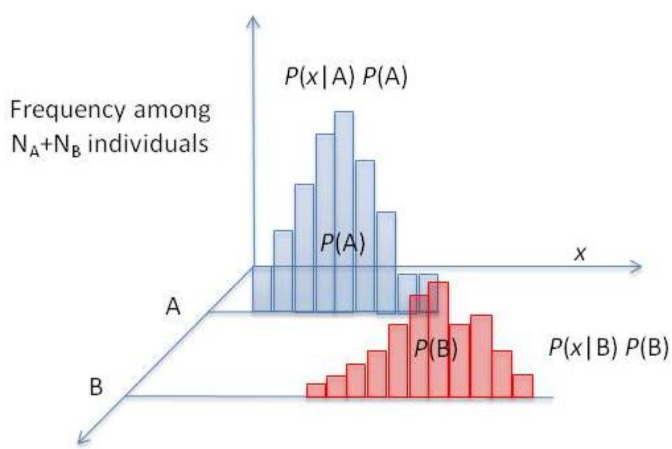

Figure 3: The histograms of Fig. 1 annotated with the meaning of each curve. The area under the blue histogram is $P(A)$, while the area under the red histogram is $P(B)$

A classifier based on Eq. 1 is called a Bayes classifier, since it is based on the Bayes formula. It could also be called a classifier based on the "relative histograms" of a training set of data, i.e. given a measurement x we label it with the name of the more numerous class of fish in the bin where x happens to fall.

If we are interested in plotting $P(A|x)$ and $P(B|x)$ we just use the Bayes expressions. The result is something like Fig. 4.

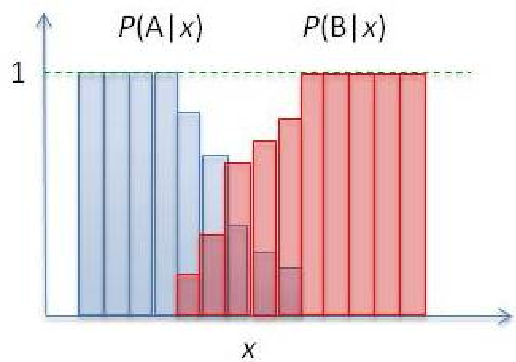

Figure 4: The probability distribution of $P(A|x)$ and $P(|x)$

If we only have the two disjoint classes A and B, then $P(x) = P(x \cap A) + P(x \cap B)$, and then

$$\begin{aligned} P(A \mid x) &= \frac{P(x \mid A)P(A)}{P(x)} \\ P(B \mid x) &= \frac{P(x \mid B)P(B)}{P(x)} \\ \Rightarrow \quad P(A \mid x) + P(B \mid x) &= \frac{P(x \cap A) + P(x \cap B)}{P(x)} = \frac{P(x)}{P(x)} = 1 \end{aligned}$$

The sum of the two curves in Fig. 4 is therefore always one. In the continuous case, the histograms for each class individually represent the probability density of each class (as in Fig. 2). The histograms relative to the total number of cases represent the products $P(x|A)P(A)$ and $P(x|B)P(B)$, the first with an integral (surface area under the curve) equal to $P(A)$, the second with a surface area $P(B)$.