Bayes (Book Chapter 2.1)

The book says that Bayesian pattern decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the trade-offs between various classification decisions using probability and the costs that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known.

That is, we assume from the start that we have some prior knowledge about the problem and relevant conditions. Bayes theorem is also important in that the Bayes classifier minimizes the probability of misclassification and thus is often said to be optimal. However, in real life we don't usually have all the information needed. If we have all the information, then Bayes classifier offers the optimal solutions.

Let's look at Bayes theorem through an example. Let's assume we want to classify two kinds of fish, Salmon w_1 and Seabass w_2. We can determine the prior probabilities $P(w_1)$ and $P(w_2)$ from the relative amounts of fish in the population or from previous catches. Prior probabilities are probabilities that we have before anything is really observed.

Now, if we try to do a classification based on prior probabilities we can decide that its w_1 if $P(w_1)>P(w_2)$, otherwise it's w_2. This is ok, if we classify just one fish. But if we have a 100 fish, it starts to sound a bit stupid. If the population has more salmons, for example, we would always decide that we caught a salmon. But all of the 100 fish cannot be salmons.

We then notice that the there is a variance in the lightness of the fish. So, we start measuring the lightness of the fish and come up with a histogram. This histogram can then be modified into a likelihood function $p(x|w_i)$. Also called class-conditional density function. This is read as "probability of w given x". This illustrates that if we got a salmon, for example, what is the probability for it to be of certain lightness.

Now, we can determine joint probability $p(x,w_i)$ which gives us a value for what's the probability for the catch to be salmon and of specific lightness.

What we eventually want to know is the posterior probability $p(w_i, x)$. This tells us that: we caught a fish with lightness x, what's the probability that it's a salmon. That is, which fish we most probably caught!

The joint probability can be written in two ways:

$$p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$$

This can then be rearranged into the Bayes formula

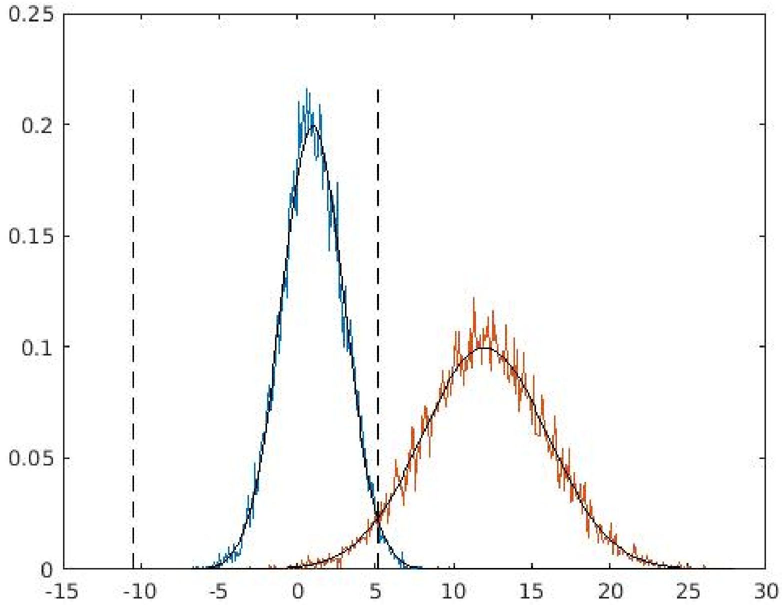
$$\boxed{P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},}$$

$$p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j).$$

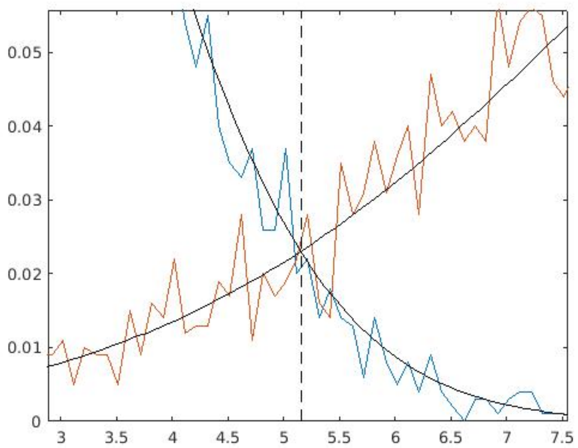
In English this could be expressed as

$$posterior = \frac{likelihood \times prior}{evidence}$$

In our Matlab programming assignment, we had two sets of data. The other one has mean of 1 and standard deviation of 2. The other has 12 and 4, respectively. We generate some random data from these distributions and draw them in a single image:

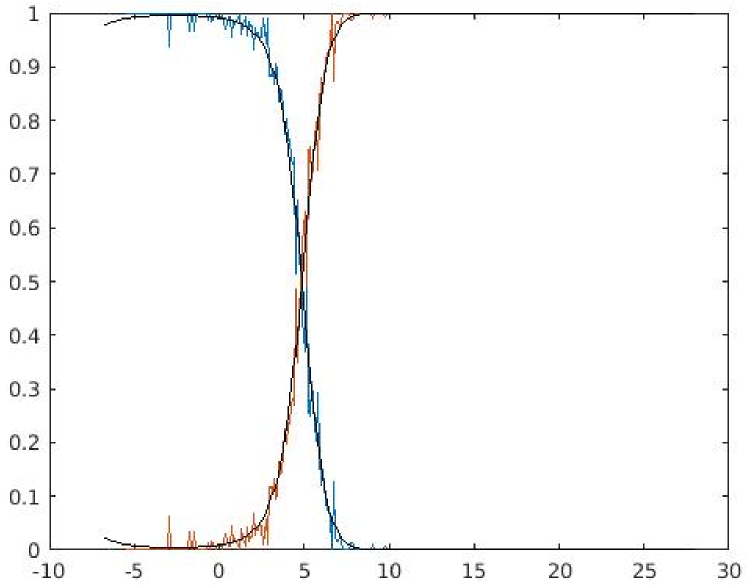


If we know that the data follows normal distribution we can use ‘solve’ to find the intersection point of the data. This intersection can later be used for classification, especially if prior probabilities are equal.

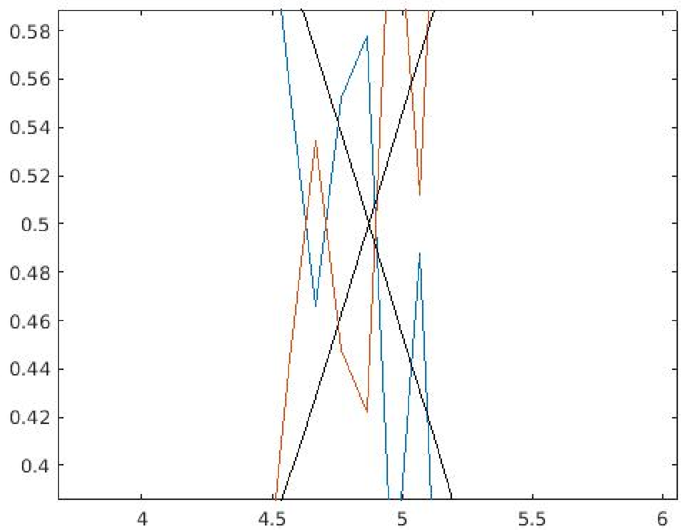


There are two intersections, around 5.1 and -10.5. The second intersection is not that clear first but if you look at the data it becomes more obvious. The other data set just diminishes a bit slower at the tails of the curve.

What if we know that the prior probabilities are 2/5 and 3/5 respectively? Let's use Bayes to calculate the posterior probabilities. The calculations should yield something like this:



From the posterior curves we can see that e.g. at 0.2 it is more probable that the observed value belongs to the first data set. However, from around 5 the situation changes and it is more probable that the value belongs to the second data set. By zooming we can see that the prior knowledge has moved the decision point closer to 4.9.



Bayesian statistician always takes into account prior knowledge. Even if the probability for the machine to lie is small, the probability for sun to go nova is even smaller.