Open your mind. LUT.
Lappeenranta University of Technology

LUT Machine Vision and Pattern Recognition 2015-11-29

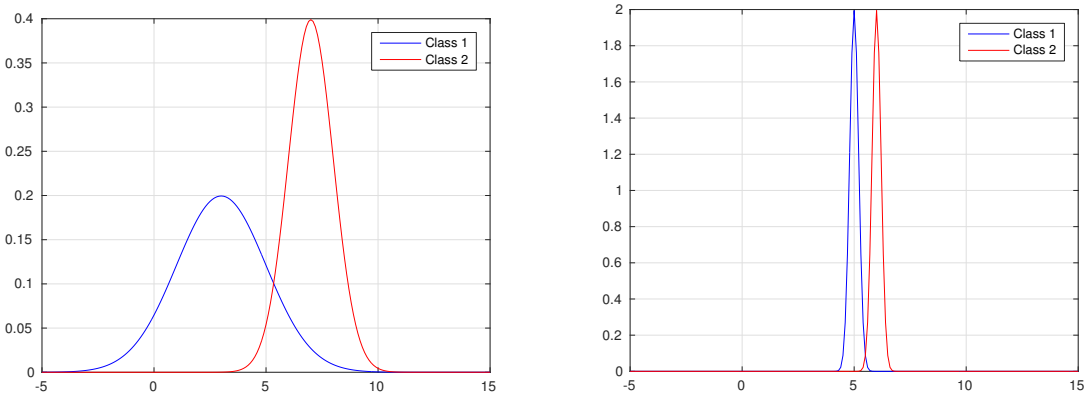
BM40A0700 Pattern Recognition

Lasse Lensu

Exercise 3 solutions: Feature processing and Bayes rule

1. Class separability (1 point): not published.

2. Separability of features (1 point): The case focusing on normally distributed features is as follows: class means for Feature1 are 3 (Class1) and 7 (Class2), and the standard deviations (STDs) are 2 (Class1) and 1 (Class2); the class means for Feature2 are 5 (Class1) and 6 (Class2), and the STDs are 0.2 for the both classes. Two Matlab plots visualizing the probability density functions (PDFs) of the features are shown in Fig. 1. The thresholds to separate the classes can be approximately



(a) (b)

Figure 1: Feature value distributions for the two classes: (a) Feature1, (b) Feature2.

determined based on the plots for the both cases.

Note: What is the underlying assumption that the threshold would be appropriate to use?

3. Feature normalisation (1 point): not published.

4. Bayes' rule (1 point): The purpose of the task was to study cancer screening with the Bayes formula and a cancer test which is not perfect: the test succeeds to reveal cancer only with probability 0.98. In addition, the test has a probability of 0.03 to show positive when the tested subject is healthy. The a priori probability of cancer in the overall population is 0.008.

   (a) What is the probability that Mr Onymous has cancer when the test is positive? Let us mark a positive result with T (true) and cancer with C (healthy subject is then a person without cancer, i.e., ¬C). Then

$$P(\mathrm{C}) = 0.008 \quad P(\neg\mathrm{C}) = 0.992 \quad P(\mathrm{T}|\mathrm{C}) = 0.98 \quad P(\mathrm{T}|\neg\mathrm{C}) = 0.03$$

   Using Bayes formula, we get the following probability for cancer after a positive test result (the law of total probability is used in the denominator to include both cases in which the

PO BOX 20, FI-53851 LAPPEENRANTA, FINLAND
E-mail: lasse.lensu@lut.fi

test can show positive):

$$P(\mathrm{C}|\mathrm{T}) = \frac{P(\mathrm{T}|\mathrm{C})P(\mathrm{C})}{P(\mathrm{T})} = \frac{P(\mathrm{T}|\mathrm{C})P(\mathrm{C})}{P(\mathrm{T}|\mathrm{C})P(\mathrm{C}) + P(\mathrm{T}|\neg\mathrm{C})P(\neg\mathrm{C})}$$
$$= \frac{0.98 \cdot 0.008}{0.98 \cdot 0.008 + 0.03 \cdot 0.992} \approx 0.209$$

The resulting probability seems very low when considering the high sensitivity of the cancer test. The reason for the low probability is the low a priori probability of cancer.

(b) What is the probability of cancer after the second test showing positive result for a specific subject? The case of two positive results is similar:

$$P(\mathrm{C}|\mathrm{TT}) = \frac{P(\mathrm{TT}|\mathrm{C})P(\mathrm{C})}{P(\mathrm{TT})} = \frac{P(\mathrm{TT}|\mathrm{C})P(\mathrm{C})}{P(\mathrm{TT}|\mathrm{C})P(\mathrm{C}) + P(\mathrm{TT}|\neg\mathrm{C})P(\neg\mathrm{C})}$$

However, what is the probability of two positive results, for example, $P(\mathrm{TT}|\mathrm{C})$? If the test results are independent, the probability of two positives for an individual subject having cancer is the product of the two individual probabilities. Then $P(\mathrm{TT}|\mathrm{C}) = P(\mathrm{T}|\mathrm{C})P(\mathrm{T}|\mathrm{C})$. Using the above formula:

$$P(\mathrm{C}|\mathrm{TT}) = \frac{0.98 \cdot 0.98 \cdot 0.008}{0.98 \cdot 0.98 \cdot 0.008 + 0.03 \cdot 0.03 \cdot 0.992} \approx 0.896$$

An alternative way to solve this would be to consider the second test case as a continuation of the first one, and use the resulting a posteriori probability of having cancer when the test shows positive the first time as the a priori probability for the second test. Intuitively this would describe a screening process with two stages, and the second stage would be performed for subjects for which the first test has shown positive. The resulting a posteriori probability of having cancer after two positive test results is identical to the previous approach.