ICT for Health
Laboratory # 6
Arrhythmia

Monica Visintin

Politecnico di Torino

2017/18

Prepare the data [1]

- From `http://archive.ics.uci.edu/ml/datasets/Arrhythmia` download the two files `arrhythmia.names` and `arrhythmia.data`
- File `arrhythmia.data` stores 280 features of 452 patients. The last column stores an integer number from 1 to 16 that specifies the patient level of cardiac arrhythmia: class 1 corresponds to absence of arrhythmia, class 16 to severe arrhythmia
- The other 279 features specify age, sex, weight, several parameters of the cardiac signals, etc

Prepare the data [2]

- We want initially to separate class 1 (healthy people) from the other classes; so we define a new class 1 (healthy) and the new class 2 (arrhythmic); we analyze the data twice
  1. to define the decision regions
  2. to measure the errors made by using the found decision regions; in particular we want to measure the probability of false and true positives and the probability of false and true negatives, possibly achieving a probability of true negatives (**test specificity**) close to one and a probability of true positives (**test sensitivity**) close to one
- Write a Python script that imports the data: note that some values are missing (shown as '?'), exclude the corresponding columns. Moreover some other columns only store zeros and they clearly carry no information: exclude also these columns.
- Then the script shall implement binary classification:
  1. Change in the matrix last column values larger than 1 with value 2 (so that we actually have only two classes)

Prepare the data [3]

2. Define vector `class_id` as the last column of matrix `arrhythmia`, define matrix `y` as the other columns
3. Define the two submatrices: `y1`, with the rows/patients corresponding to `class_id=1`, and `y2`, with the rows/patients corresponding to `class_id=2`
4. Find `x1`, the mean of the row vectors in `y1`, and `x2`, the mean of the row vectors in `y2`; then `x1` and `x2` are the representative vectors of classes 1 and 2, respectively; they are rows with as many elements as the number of remaining features.
5. Apply the **minimum distance criterion** to associate each row of `y` with either `est_class_id=1` or `est_class_id=2`.
6. Measure the probabilities of true/false positives and the probabilities of true/false negatives
7. Use the **Bayes criterion**. We want to consider the more general case in which

$$\mathcal{H}_1: \quad \mathbf{y} = \mathbf{x}_1 + \boldsymbol{\nu}_1, \quad \boldsymbol{\nu}_1 \in \mathcal{N}(0, \mathbf{R}_1)$$

$$\mathcal{H}_2: \quad \mathbf{y} = \mathbf{x}_2 + \boldsymbol{\nu}_2, \quad \boldsymbol{\nu}_2 \in \mathcal{N}(0, \mathbf{R}_2)$$

and $\pi_1 = P(\mathcal{H}_1) \neq \pi_2 = P(\mathcal{H}_2)$.

Prepare the data [4]

8. Measure π_1 (number of patients without arrhythmia over the total number of patients in the set) and π_2 (number of patients with arrhythmia over the total number of patients in the set)
9. Measure the covariance matrice $\mathbf{R}_1$ of `y1`; diagonalize the matrix: $\mathbf{R}_1 = \mathbf{U}_1 \boldsymbol{\Lambda}_1 \mathbf{U}_1^T$
10. Measure the covariance matrice $\mathbf{R}_2$ of `y2`; diagonalize the matrix: $\mathbf{R}_2 = \mathbf{U}_2 \boldsymbol{\Lambda}_2 \mathbf{U}_2^T$
11. At this point, the assumed probability density function of $\mathbf{y}$ given $\mathcal{H} = 1$ (i.e. healthy patient) is

$$f_{\mathbf{y}|\mathcal{H}_1}(\mathbf{u}) = \frac{1}{\sqrt{(2\pi)^F \det(\mathbf{R}_1)}} \exp\left(\frac{-1}{2}(\mathbf{u}-\mathbf{x}_1)^T \mathbf{R}_1^{-1}(\mathbf{u}-\mathbf{x}_1)\right)$$

whereas the probability density function of $\mathbf{y}$ given $\mathcal{H} = 2$ (i.e. arrhythmic patient) is

$$f_{\mathbf{y}|\mathcal{H}_2}(\mathbf{u}) = \frac{1}{\sqrt{(2\pi)^F \det(\mathbf{R}_2)}} \exp\left(\frac{-1}{2}(\mathbf{u}-\mathbf{x}_2)^T \mathbf{R}_2^{-1}(\mathbf{u}-\mathbf{x}_2)\right)$$

Prepare the data [5]

However $\mathbf{R}_1^{-1}$ and $\mathbf{R}_2^{-1}$ cannot be evaluated, because some of the eigenvalues are very close to zero.

12. Keep only F' (to be optimized) columns of $\mathbf{U}_2$, corresponding to the largest eigenvalues, getting the matrix `UF2`, and keep only F' columns of $\mathbf{U}_1$, corresponding to the largest eigenvalues, getting the matrix `UF1` (note that you could also consider the case in which you keep F_1 columns of $\mathbf{R}_1$ and F_2 columns of $\mathbf{R}_2$, with $F_1 \neq F_2$, note that we are performing PCA, Principal Component Analysis)
13. project `y1` (F columns) onto `UF1` and get the new matrix `z1` (only F' columns); the covariance matrices of `z1` is now diagonal (check it); project `y2` (F columns) onto `UF2` and get the new matrix `z2` (only F' columns);
14. find the means of `z1` (those rows of z corresponding to `class_id=1`) and `z2` (those rows of z corresponding to `class_id=2`) and call them `w1` and `w2` (each of these vectors has only F' elements)
15. write the pdf's of `z1` $f_{\mathbf{z}|\mathcal{H}_1}(\mathbf{u})$ and `z2` $f_{\mathbf{z}|\mathcal{H}_2}(\mathbf{u})$
16. project matrix `y` onto `UF1` to get matrix `s1` and onto `UF2` to get matrix `s2`

Prepare the data [6]

17. According to the MAP/Bayes criterion, compare the probabilities $\pi_1 f_{\mathbf{z}|\mathcal{H}_1}(s_1(n))$ and $\pi_2 f_{\mathbf{z}|\mathcal{H}_2}(s_2(n))$ to estimate the class (i.e. `est_class_id`)
18. measure the probabilities of true/false positives and the probabilities of true/false negatives

- Repeat the exercise using now the 16 classes and the minimum distance rule (note that classes 11, 12 and 13 never occur in the data set). Note that the sensitivity and specificity can only be used in binary classification. In case of multiclass classification, the **confusion matrix** is measured: the element in position i, j of the matrix is the probability that the estimated class is j given that the true class is i (the sum of the elements in a row must be 1).
- Write the report