Pattern Recognition
Department of Electrical and Information Engineering
University of Cassino and Southern Latium, Second Semester 2018
Homework Assignment 1
Assigned 10 April 2018; due 11:59pm, 27 April 2018

Problem 1.1 [5%] Describe (max one page) an application of pattern recognition related to your research. What are the features? What is the decision to be made? Comment on how one might solve the problem.

Problem 1.2 [15%]. In a particular two–class problem, the conditional densities for a scalar feature x are $p(x|\omega_1) = k_1 \cdot exp\left(-\frac{x^2}{15}\right)$ and $p(x|\omega_2) = k_2 \cdot exp\left(-\frac{(x-7)^2}{14}\right)$

(a) Find k_1 and k_2, and plot the two densities on a single graph using `Matlab`.

(b) Find the decision regions which minimize the average probability of error and indicate them on the plot you made in part (a) for the following cases:

(b.1) $\frac{P(\omega_2)}{P(\omega_1)} = 1$

(b.2) $\frac{P(\omega_2)}{P(\omega_1)} = 0.5$

(b.3) $\frac{P(\omega_2)}{P(\omega_1)} = 4$

Problem 1.3 [20%] Consider a medical diagnosis problem where a biochemical test is used for screening patients. The test returns a result close to 0 for healthy patients and close to 1 for sick patients, according to the following likelihood functions: $p(x|\omega_1) = N(0.0, 0.1)$ and $p(x|\omega_2) = N(1.5, 0.3)$ where $N(\mu, \sigma)$ is the Gaussian univariate density, ω_1 is the class of the healthy patients and ω_2 is the class of the sick patients. Assume that, on average, 1 out of 10,000 patients is sick, and the following costs:

1. $\lambda_{21} = 1,500$ euros for improductive test cost in the case of wrongly diagnosed disease;
2. $\lambda_{12} = 800,000$ euros for reimbursement to the patient in the case of not diagnosed disease;
3. $\lambda_{11} = \lambda_{22} = 0$ in the case of correct diagnosis

and define the decision rule that minimizes the Conditional Risk.

Note that $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ is the cost produced by a decision for class ω_i when the true class is ω_j.

Problem 1.4 [30%] The file `hw1data.zip` available in the homework folder contains samples coming from a two–class problem, each made of 10 numerical features and a binary label (± 1). Split the data into training and test sets by randomly selecting 25% of the examples from each class for the test set.

(a) Using the training data, implement a linear classifier and a quadratic classifier and evaluate on the test set the True Positive Rate obtained for a False Positive Rate of 0.1. Repeat the above steps several times. What is the average obtained TPR for each of the implemented classifiers? Discuss your results.

(b) Repeat the previous exercise, but this time implementing a $k - NN$ classifier. Experiment with several values for k. Discuss your results and compare them with those obtained in the previous exercise.

Problem 1.5 [30%] Use the dataset of Problem 1.4 and, using a model among the ones you know (linear, quadratic, $k - NN$), build the classifier with the best accuracy on that classification problem. To this aim, you can consider several training/test splits with different percentages for training and test sets. Prepare a `Matlab` fuction called test.m (`function y=test(A)`) that implements your best choice classifier. The function will accept a matrix `A` and return a vector `y` having the same number of rows in `A`. The matrix `A` will contain several samples (one for each row) organized in the same way described in Problem 1.4, but without label (i.e. each row will contain containing a sample with 10 numerical features). The function will classify each of the samples `A(i,:)`, providing the predicted class in `y(i)`. Your function must be submitted and will be run on a separate matrix containing new test data. Your grade will be based on the performance of your classifier on the new test data, which will contain a very large number of examples generated from the same distribution.