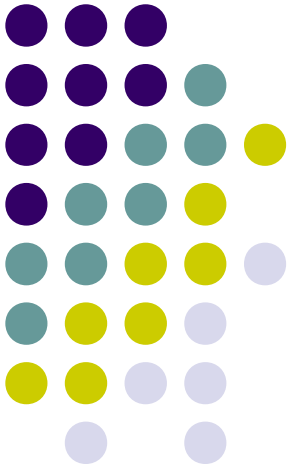
Pattern Recognition

Lab3:
K-NN classifier
ROC curve

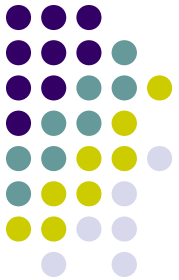
Francesco Tortorella

University of Cassino and Southern Latium

Cassino, Italy



Lab 3.1

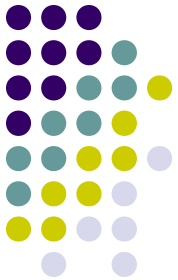


- Consider two classes with Gaussian densities:

$$p(x|\omega_1) = N(\mu_1, \Sigma)$$
$$p(x|\omega_2) = N(\mu_2, \Sigma)$$

- Consider the prior probability P_1 varying from 0.1 to 0.9 with step 0.05. For each value
  - Extract two sets (TrainSet and TestSet) with number of samples proportional to the values of the priors
  - Build the linear classifier and measure the error rate on the TestSet
  - Using TrainSet, build a k-NN (with k varying from 1 to 21, step 2) and measure the error rate on the TestSet
  - Plot the error rate vs. P_1 for the linear classifier
  - For each k, plot the error rate vs. P_1

Lab 3.2

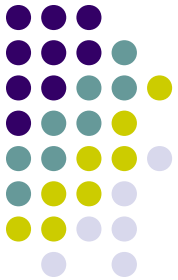


- Do the same tasks required by 2.1 on two classes with Gaussian densities :

$$p(x|\omega_1) = N(\mu_1, \Sigma_1) \qquad \Sigma_1 \neq \Sigma_2$$
$$p(x|\omega_2) = N(\mu_2, \Sigma_2)$$

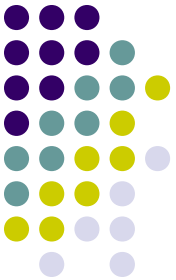
- Consider the prior probability P_1 varying from 0.1 to 0.9 with step 0.05. For each value
  - Extract two sets (TrainSet and TestSet) with number of samples reflecting the values of the priors
  - Build the linear classifier and measure the error rate on the TestSet
  - Build the quadratic classifier and measure the error rate on the TestSet
  - Using TrainSet, build a k-NN (with k varying from 1 to 21, step 2) and measure the error rate on the TestSet
  - Plot the error rate vs. P_1 for the linear classifier and for the quadratic classifier
  - For each k, plot the error rate vs. P_1

Lab 3.3



- Read the file ‘pima-indians-diabetes.data’
- Part the set into two equal subsets: assume one (PimaTr) as training set and the other one (PimaTest) as test set
- Starting from the training set, build a linear classifier, a quadratic classifier and a k-NN with k varying from 1 to 15, step 2.
- For each of the classifiers built, estimate the ROC curve on the test set.

Lab 3.4



- Read the file ‘pima-indians-diabetes.data’
- Part the set into two subsets: assume one (PimaTr) as training set and the other one (PimaTest) as test set
- Starting from the training set, build a linear classifier, a quadratic classifier and a k-NN with k varying from 1 to 15, step 2.
- Consider different sizes of the training set (25%, 50%, 75%) and analyze how the AUC changes.