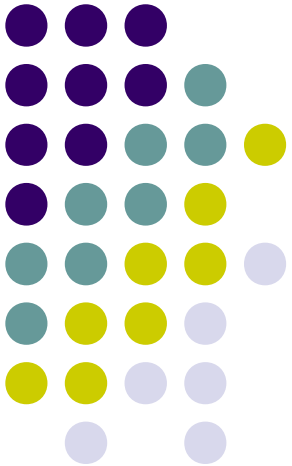
Pattern Recognition

Lab2:
Linear & Quadratic Classifiers

Francesco Tortorella

University of Cassino and Southern Latium

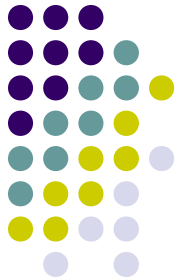
Cassino, Italy



Lab 2

- Real medical problem: The Pima Indians in the U.S. have the highest rates of diabetes and obesity in the United States. They reside mainly in the desert regions of Arizona and have the world's highest recorded prevalence (about 38%) and incidence of type 2 diabetes.
- The genetically similar Pima Indians in Mexico have a prevalence of about 6.9%.

Lab 2



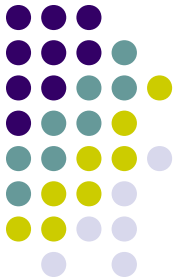
- Real data set ‘Pima.data’ (8 features, 2 classes)
  - 1. Number of times pregnant
  - 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
  - 3. Diastolic blood pressure (mm Hg)
  - 4. Triceps skin fold thickness (mm)
  - 5. 2-Hour serum insulin (mu U/ml)
  - 6. Body mass index (weight in kg/(height in m)^2)
  - 7. Diabetes pedigree function
  - 8. Age (years)
  - 9. Class variable (0 or 1)

Lab 2

- Class Distribution: (class value 1 is interpreted as "tested positive for diabetes")
  - Class Value Number of instances
    0 500
    1 268
- Brief statistical analysis:

Attribute number:	Mean:	Standard Deviation:
1.	3.8	3.4
2.	120.9	32.0
3.	69.1	19.4
4.	20.5	16.0
5.	79.8	115.2
6.	32.0	7.9
7.	0.5	0.3
8.	33.2	11.8

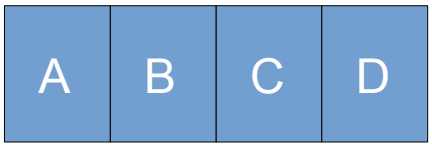
Lab 2.1



- 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 (Take care to preserve the original priors!).
- Starting from the training set, build a linear classifier and a quadratic classifier and compare the accuracies on the test set
- Consider different sizes of the training set (25%, 50%, 75%) and analyze how the accuracy changes.

Lab 2.2

- The results you obtained in 2.1 are for a particular partition of the original data set into a training and a test set. What if the data set is split in a different way?
- Consider the original data set divided into 4 equal parts (A,B,C,D)



- Let us assume that the previous experiment with a training set containing the 75% of the samples was made considering $A \cup B \cup C$ as training set and D as test set. Repeat this experiment, but using in turn, as test set, A, B and C (i.e. if B is used as test set, the training set is $A \cup C \cup D$) and analyze how the accuracy changes.