**Project 2 Outline (completed from previous lecture notes 02.19.2015.Tatarin.draft2 ) :**

Using your 2 data sets from project 1 for which you performed LDA, now perform Support Vector Machine (SVM) classification between the same two classes . We are interested in improving our LDA results with SVM. If your LDA results for a data set exceeded 90%, choose either two new classes or if you prefer, another data set for analysis

if you change datasets , insert again a description of attributes and classes , but you do not need to present again the histograms

as for the visualization of the data

project data in dimension 3

either by selecting three attributes with largest t-test values ,

or more efficiently by applying classical PCA to all attributes and keeping the three eigenvectors with largest eigenvalues

SVM Classification **for 2 classes only**

Suggested Software Packages

SVM Light, e1071, and many others that can befound online

Website Suggestions for SVM Light:

R: http://www.inside-r.org/packages/cran/klaR/docs/svmlight

Matlab: http://mex-svm.sourceforge.net/

Mathematica: http://malchiodi.di.unimi.it/software/svMathematica/

• Use different kernels:

**–** Polynomial kernels of degree {2, 3,4} (try all 3)

**–** Gaussian kernels: *K*(*x*, *y*) = exp(-||x-y||2/ σ2)

for this kernel try multiple values of scale parameter σ and report corresponding classification performances to select the best σ You can for instance explore the range of σvalues by successive dichotomies; but some SVM softwares automatically optimize on sigma

• maximum margin optimization performed by SVM softwares depends on the choice of a parameter $C>0$ which is a weight coefficient for the sum of squared sizes of classification errors; many svm softwares automatically optimize the choice of C by default, although the user can preassign any value of C

display this optimal performance, but also compute and display the performances obtained for 4 different values of C

• Construct a 2x2 "Confusion (or Error)Matrix" (p**ij**) for the training set and another

confusion matrix for the test set. Compute the standard errors on each p**ij**

as indicated in the lectures

• Compare the two matrices to evaluate the generalization capacity of the SVM classifier

• For each kernel type, after computing the best SVM of that type , give the explicit equation of the Separator: *Sep*(*x*) (see lecture notes 02-19-2015 draft2)

• Call α**i** the non zero coefficients associated to the S support vectors x**i** . Give the list of these coefficients ranked in decreasing order of their absolute values and plot these decreasing absolute values on a graph.

Display the support vectors in your 3D visualization, display specifically the 5 support vectors with the highest | α**i** | ; look at these cases and try to see why they are hard to classify;

As explained in lectures, call Xi the vector in Hilbert space H associated to the true support vector x**i** and compute the distances (in Hilbert space H) Dij = || Xi - Xj ||H

Identify the two support vectors z**i** and z**j** which realize the minimum of Dij and check what happens to the separator formula if you replace this specific z**i** by z**j**

• Toy Problem: To be included in project 2

**(see details in** lecture notes 02.19.2015.Tatarin.draft2)