****

**Pattern Recognition**

**F. Meriaudeau**

**HW5 – Feature selection**

**Note:** Though some of these exercises specifically require MATLAB, you may use MATLAB to help with plotting, computation, and/or verification of your results. This is in fact encouraged.

**1. Hypothesis testing : the *t-*test**

The first test in feature selection is too look at each feature *individually* and check whether or not it is an informative one. If not, the feature is discarded. To this end, statistical tests are commonly used. The idea is to test whether the mean values that a feature has in two class differ *significantly*. Assuming that the data in the classes are normally distributed, the so-called *t-*test is a popular choice.

The goal of the t-test is to determine which of the following Hypothesis is true:

H1: the mean values of the feature in the two classes are different (***alternative hypothesis***)

H0: the mean values of the feature in the two class are equal (***null hypothesis***)

If the null hypothesis holds true, the feature is discarded. If the alternative hypothesis holds true, the feature is selected. The hypothesis test is carried out against the so-called **significance level**, ****, which corresponds to the probability of committing an error in our decision. Typical values are =0,05 and = 0,001.

**Example** : Assume that a feature follows Gaussian distributions in both classes of a 2-class classification problem. The respective means are m1=8,75 and m2=9; their common variance is 2=4.

1. Generate the vector x1 and x2, each containing N=1000 samples from the first and second distribution, respectively.
2. Pretend the means and variance are unknown. Assumed to be known are x1 and x2 and the fact that they come from distributions with equal (yet unknown) variance. Use the t-test to check whether the mean values of the two distributions differ significantly using as a significance level ,  = 0,05.
3. Repeat with  = 0,001 and draw conclusions.

Use ttest2 function provided.

**2. Fisher discriminant ratio (FDR).**

FDR is commonly employed to quantify the discriminatory power of individual features between two equiprobable classes. In other word, it is independent of the type of the classe distribution. Let m1 and m2 be the respective mean values and 12 and 22 the respective variances. The FDR is defined as:

Example: Generate N=500 5-dimensional data vectors that stem from two equiprobable classes w1 and w2. The classes are modeled by Gaussian distributions with respectively, means m1= [1,1,0,6,3] and m2= [11.5,11,10,6.5,4]T and covariance matrices

and

Compute de FDR values for the five features and draw conclusions.

Use Fisher function provided.

Example 2:

The goal of this example is to demonstrate the use of the FDR in ranking a number of features with respect to their class discriminatory power.

Features (Mean,std,skewness and kurtosis) are extracted from various ROIs of liver disease (cirrhotic or fatty infiltration)

Using the fattyliver.dat and cirrhoticliver.dat compute the FDR and rank the features.

**3. Measures based on Scatter Matrices.**

Scatter matrices are among the most popular measures for quantifying the way feature vectors “scatter” in the feature space. Because of their rich physical meaning, a number of class-separability measures are built around them. Three such measures are the following:

Sm is the mixture matrix, Sw is the within class matrix and Sb is the between class matrix.

Use the J3 criterion to choose the best l features out of m (m>l) for the data of example 2 section 2. To do this

* Normalize the value of each feature to have zero mean and unit variance (use normaliseStd)

Select three out of the four features that result in the best J3 value (Use the ScatterMatrices function)

* Plot the data points in the 3-dimensionnal space for the best feature combination

**4.Feature subset selection**

We will investigate the suboptimal searching techniques/the exhaustive search.

Assume we are given two images of different textures. We associate each texture with a different class. In each of the two images, we form 25 ROIs. These 50 total patterns form the training set, which will be used for selecting the “best” combination of features.

Thus for each pattern 20 features are generated (four first order statistics and 16 (4 directions (0°,90°,45° and 135° \* 4 features( contrast, correlation, energy and homogeneity) features derived from GLCM.

The features are ordered (4 stat, 4 contrats,….).

Once the feature generation is completed, two 25x20 arrays are formed dataclass1.dat and dataclass2.dat

We assume that we are working in a 2-dimensionnal feature space and we evaluate the performance of exhaustivesearch, SFS, SBS,SFFS as follows:

* Normalize the value of each feature to have zero mean and unit variance.
* Rank them utilizing scalar feature selection, which employs the FDR criterion and a cross-correlation measure between pairs of features. Set a1=0.2 and a2=0.8 (use ScalarfeatureSelectionRanking)
* Select the 14 highest-ranked features and employ the exhaustive search method with J3 criterion to select the best combination of two features.
* Repeat the last step for SFS, SBS and SFFS and comment the results