Técnicas de Reconhecimento de Padrões
Pattern Recognition Techniques

2009/2010

Practical Exercises : *Class # 0*

Data Sets from: "Pattern Recognition: Concepts, Methods and Applications"

Topics

Pattern Discrimination

- Decision Regions and Functions
- Hyperplane Separability
- Feature Space Metrics

Linear Discriminant Classifiers

1. A classifier uses the following linear discriminant in a 3-dimensional space:

$$d(\mathbf{x}) = x_1 + 2x_2 + x_3 + 1$$

   a) Compute the distance of the discriminant from the origin.

   b) Give an example of a pattern whose classification is borderline ($d(\mathbf{x}) = 0$)

   c) Compute the distance of the feature vector $[1 \ -1 \ 0]'$ from the discriminant.

2. A pattern classification problem has one feature x and two classes of points $\omega_1 = \{-2, 1, 1.5\}; \omega_2 = \{-1, -0.5, 0\}$.

   a) Determine a linear discriminant of the two classes in a two dimensional space using features $y_1 = x$ and $y_2 = x^2$ and write a linear decision rule.

   b) Determine the quadratic classifier in the original feature space that corresponds to the previous linear classifier.

3. Consider a two class one-dimensional problem with one feature x and classes $\omega_1 = \{-1, 1\}; \omega_2 = \{2, 0\}$.

   a) Show that the classes are linearly separable in a two dimensional space with feature vectors $\mathbf{y} = [x^2 \ x^3]'$ and write a linear decision rule.

b) Using the previous linear decision rule write the corresponding rule in the original feature space.

4. Consider the equidistant surfaces relative to two prototype patterns, using the city-block and Chebychev metrics in a two-dimensional feature space as shown in the book Figure 2-10. In which cases do the decision functions correspond to straight lines?

5. Which of the following matrices can be used in a 2-dimensional linear transformation preserving the Mahalanobis distance?

a)
$$A = \begin{bmatrix} 2 & 1 \\ -1 & 1 \end{bmatrix}$$

b)
$$B = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

c)
$$C = \begin{bmatrix} 1 & 0.5 \\ 0.5 & -1 \end{bmatrix}$$

Explain why.

Técnicas de Reconhecimento de Padrões
Pattern Recognition Techniques

2009/2010

Practical Exercises : *Class # 1*

Data Sets from:
"Pattern Recognition: Concepts, Methods and Applications"

Topics

PCA (dimension reduction)

Kruskal Wallis Test

Feature Assessement

Exercise 1 - PCA -Dimension Reduction

Consider the `cork_stoppers.xls` data set containing measurements performed automatically by an image processing system on 150 cork stoppers belonging to three classes (ω_1 - Super, ω_2 - Average and ω_3 - Poor)

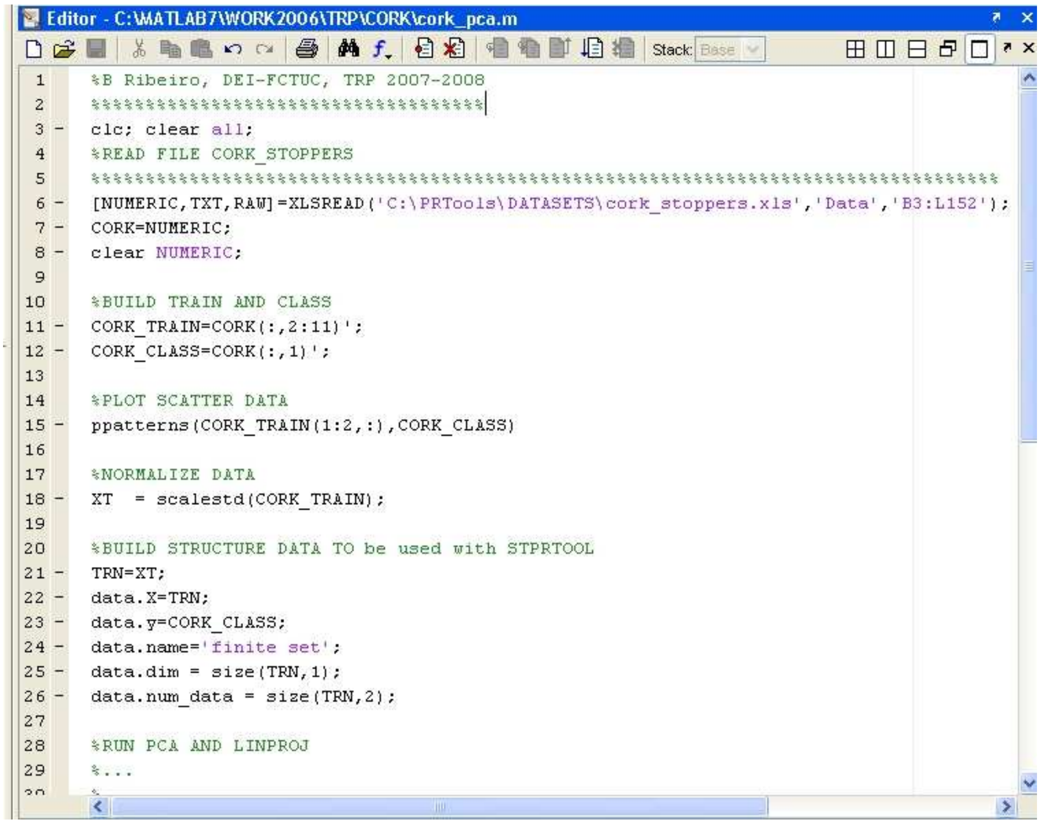
Use the `m` files (`pca.m` and `linproj.m`) available on the "STPRTool - Statistical Pattern Recognition Toolbox for Matlab".

1. Write the `function scalestd` to normalize your data features (mean 0 and standard deviation 1); see the book section "normalizing data"
2. Find the Principal Components in the `cork_stoppers.xls` data set.
3. Plot the first two components in a scatter diagram for the first two classes in the data set (ω_1 and ω_2) ; use the `ppatterns` function available in STPRTool
4. Plot the eigenvalues

Hints

1. Build a `trpstartup.m` file to put the "STPRTool - Statistical Pattern Recognition Toolbox for Matlab" on your path. The file should contain the instruction:

```
stprpath('C:\MATLAB7\toolbox:\stprtool');
```




```
%B Ribeiro, DEI-FCTUC, TRP 2007-2008
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clc; clear all;
%READ FILE CORK_STOPPERS
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[NUMERIC,TXT,RAW]=XLSREAD('C:\PRTools\DATASETS\cork_stoppers.xls','Data','B3:L152');
CORK=NUMERIC;
clear NUMERIC;

%BUILD TRAIN AND CLASS
CORK_TRAIN=CORK(:,2:11)';
CORK_CLASS=CORK(:,1)';

%PLOT SCATTER DATA
ppatterns(CORK_TRAIN(1:2,:),CORK_CLASS)

%NORMALIZE DATA
XT  = scalestd(CORK_TRAIN);

%BUILD STRUCTURE DATA TO be used with STPRTOOL
TRN=XT;
data.X=TRN;
data.y=CORK_CLASS;
data.name='finite set';
data.dim = size(TRN,1);
data.num_data = size(TRN,2);

%RUN PCA AND LINPROJ
%...
```

Figura 1: Constructing Data Training and Test sets.

2. You should run the `trpstartup` command to initiate your session and be able to access all the programs available in the STPRTool
3. Use the following lines of code to read the Excel file `cork_stoppers.xls` into Matlab

Exercise 2 - Kruskal Wallis Test

Cardiotocography is a popular diagnostic method in Obstetrics. Consider the CTG data set `CTG.xls` containing measurements and classification results of cardiographic (CTG) examinations of 2126 foetuses.

1. Perform Kruskal-Wallis tests for all the features and sort them by decreasing discriminative capacity (10 classes).
2. Using the rank values of the tests and box plots determine the contributions for class discrimination of the best three features.

Hints

1. Use Matlab function `kruskalwallis`

2. Write in your code the following instruction

```
[p,table,stats] = kruskalwallis(CTG_DATA)
```

3. You should be able to see the following Box Plot Figure:

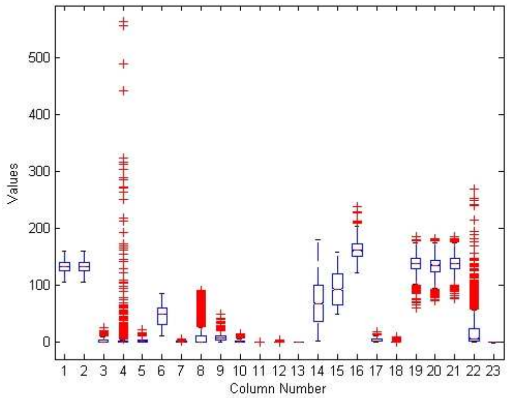


Figura 2: Box Plot in CTG data set

Exercise3 - Feature Assessement

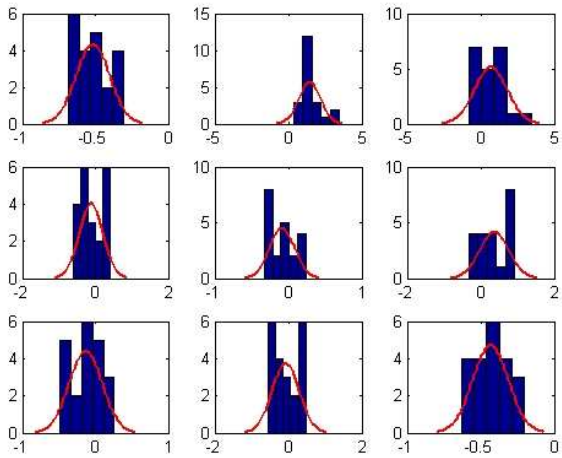


Figura 3: Features distribution in the Breast Tissue data set

Consider the `Breast Tissue.xls` data set which contains 106 electrical impedance measurements performed on samples of breast tissue. Six classes were identified:

car - carcinoma (21 cases)

fad - fibro-adenoma (15 cases)

mas - masthopaty (18 cases)

gla - glandular (16 cases)

con - connective (14 cases)

adi - adipose (22 cases)

Determine for each pattern class, which features distribution can be reasonably described by the normal model.

1. Normalize your data using `function scalestd`
2. Use Matlab functions `hist.m` and `histfit.m`
3. Use subplot to get the following graph

Técnicas de Reconhecimento de Padrões
Pattern Recognition Techniques

2009/2010

Practical Exercises : *Class # 2*

Data Sets from:
"Pattern Recognition: Concepts, Methods and Applications"

Topics

Statistical Pattern Recognition

- Linear Discriminant Classifiers
  - Euclidian Linear Discriminat
  - Mahalanobis Linear Discriminant
- Fisher Linear Discriminant
- Perceptron and MultiPerceptron

Linear Discriminant Classifiers

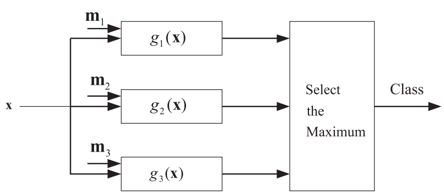


Figura 1: Discriminant functions

Euclidian Linear Discriminants

$$
\begin{aligned}
g_k(\mathbf{x}) &= \mathbf{w}_k'\mathbf{x} + \mathbf{w}_{k,0} \qquad (1)\\
\text{with } \mathbf{w}_k &= \mathbf{m}_k\\
\text{and } \mathbf{w}_{k,0} &= -0.5||\mathbf{m}_k||^2
\end{aligned}
$$

Mahalanobis Linear Discriminants

$$
\begin{aligned}
g_k(\mathbf{x}) &= \mathbf{w}_k'\mathbf{x} + \mathbf{w}_{k,0} \\
\text{with } \mathbf{w}_k &= \mathbf{C}^{-1}\mathbf{m}_k \\
\text{and } \mathbf{w}_{k,0} &= -0.5\mathbf{m}_k'\mathbf{C}^{-1}\mathbf{m}_k
\end{aligned}
\tag{2}
$$

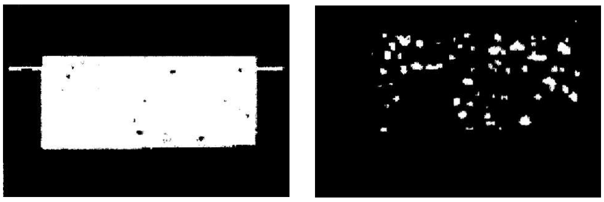


Figura 2: Cork Stoppers: image (left) and the corresponding binary image (right)

Consider the `cork_stoppers.xls` data set containing measurements performed automatically by an image processing system on 150 cork stoppers (see Figure 2) belonging to three classes (ω_1 - Super, ω_2 - Average and ω_3 - Poor).

In your exercise consider two features (Feature 1-ART and Feature 3-PRT). Consider only the classes ω_1 and ω_2.

1. Plot Histograms for the two classes (take Feature 2-N only for simplicity) as shown in Figure 3. You can use the following code:

```
%PLOT HISTOGRAMS DATA CHOOSING N - feature
hist(CORK(1:50,3),14);
hold on
h = findobj(gca,'Type','patch');
set(h,'FaceColor','r','EdgeColor','w')
hist(CORK(51:100,3), 14);
```

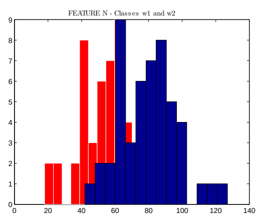


Figura 3: Histogram: Feature N for two classes ω_1 and ω_2

2. Find the prototypes $\mathbf{m_1}$ and $\mathbf{m_2}$ for the first two classes. Calculate the Covariance matrices for classes ω_1 and ω_2. Calculate the Pooled Covariance Matrix $\mathbf{C}$ and its inverse $\mathbf{C}^{-1}$.

3. Use equations above for Linear Discriminants Analysis (LDA).Evaluate the vectors predict_Euclidian and predict_Mahalanobis vector containing the predicted classes corresponding to the methods

'Euclidian Linear Discriminant',

'Mahalanobis Linear Discriminant'.

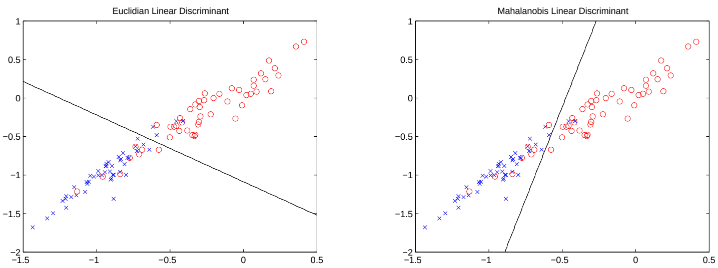
4. Evaluate the classification results and error using both discriminant functions using the STPRTool function cerror. Example: cerror(predict,trn.y).
   You should check the following values:
   Euclidian LDA 13% Error $\longrightarrow$ 87% Performance
   Mahalanobis LDA 10% Error $\longrightarrow$ 90% Performance

5. Visualization of the Linear Discriminants

   (a) You have to build each model so that the visualization can be easy using functions of STPRTool ppatterns and pboundary

   (b) Example:

      – modelEuclidian.W = wk; % assume wk is the data structure in MatLab with vector weights
      – modelEuclidian.b = wk0; % assume wko is the data structure bias values
      – modelEuclidian.fun ='linclass'; % the linear classifier
      – model.Euclidian.name ='Euclidian Linear Discriminant'



FISHER Linear Discriminant, Perceptron and MultiPerceptron

Use the MatLab functions available in STRPTool: `fld`, `fldqp`, `perceptron` and `mperceptron` for classes (ω_1 - Super, ω_2 - Average) in order to evaluate the performance of the following classifiers:

(a) Perceptron: One Feature 2-N

(b) Perceptron: Two Features: 1-ART and 2-N

(c) MultiPerceptron: T wo Features: 1-ART and 2-N; 3 classes $\omega_i, i = 1, \cdots 3$

(d) Fisher:Two Features: 2-N and 3-PRT $\mapsto$ `fld`

(e) Fisher:Two Features: 2-N and 3-PRT (Quadratic Programming ↦`fldqp`)

(f) K-Nearest Neighbor (KNN): Two Features: 2-N and 3-PRT

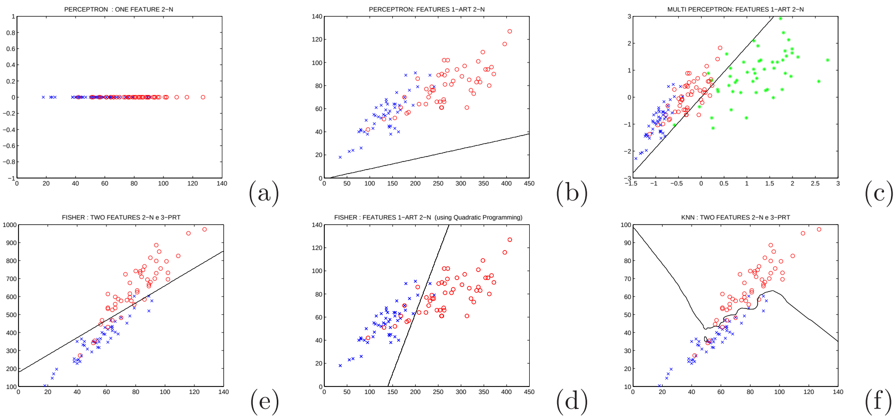


Figura 4: Linear Discriminants in CORK Data Set

1. Generate a data test set `CORK_TST` extracting 50 samples `CORK_TRAIN` by using random indexes
   Example:
   ```
   % BUILDING A TEST SET USING RANDOMLY CHOSEN EXAMPLES
   a=1;b=150;
   idx=floor( a + (b-a) * rand(50,1));
   CORK_TST = CORK_TRAIN(:,idx);
   ```

2. Plot results as shown in Figure 4 corresponding to:

   Perceptron

   MultiPerceptron

   Fisher Linear Discriminant

   K-Nearest Neighbor (KNN)
   Use `ppatterns` and `pboundary` functions

3. Evaluate prediction errors in Train and Test data set. Setup `options.tmax = threshold, threshold e.g.` 2000 in case you use `perceptron`; otherwise it might not converge.
   (Hint: Use function `cerror`)

4. Compare Classifiers (Hint: Check Model Parameters (W and b returned by tested methods)

Handwritten Recognition: UCI Machine Learning DataSet

You can access the link http://archive.ics.uci.edu/ml/ and choose to download the Semeion Handwritten Digit Data Set. In this problem 1593 handwritten digits from around 80 persons were scanned, stretched in a rectangular box 16×16 in a gray scale of 256 values. The objective is to find a classifier to separate the digits whose features are obtained from binarized images.

The first 36 images of the data set are represented in Figure 5

1. Download: Data Folder, Data Set Description
2. Build an Excel file with two sheets 'Semeion' and 'description'
3. Using Matlab and the PR Toolbox elaborate a program to
   (a) read raw data,
   (b) visualize data; in your figure use the following MatLab command
   `title 'Raw SEMEION Data Visualization (after import from Excel)'`
   (c) Build Train and Test Data Sets. Use `TRAIN_SEMEION` and `TEST_SEMEION`
4. Use PCA component Analysis to perform Feature extraction
5. Visualize data; use the following MatLab command
   `title 'SEMEION Data Visualization after PCA Projection'`
6. Perform a study to evaluate the variance of the data embedded on the extracted components. (e.g. 2, 10, 50 components);
7. Build Linear Classifiers to deal with this multiclass problem (e.g FLD, Perceptron, MultiPerceptron)
8. Evaluate Classifier Performance by computing the error in the train and test data set using `cerror`
9. Visualize boundary decision results using `ppatterns` and `pboundary`

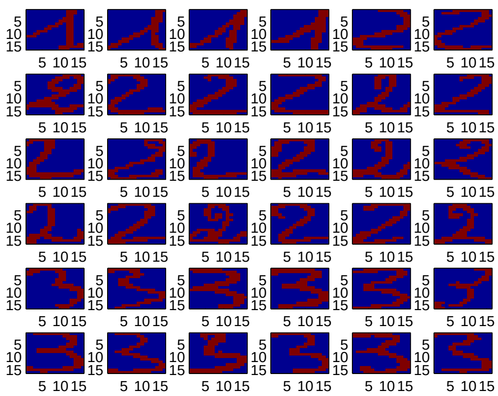


Figura 5: 36 images of SEMEION data set

Técnicas de Reconhecimento de Padrões
Pattern Recognition Techniques

2009/2010

Practical Exercises : *Class # 3*

Data Sets from:
"Pattern Recognition: Concepts, Methods and Applications"

Topics

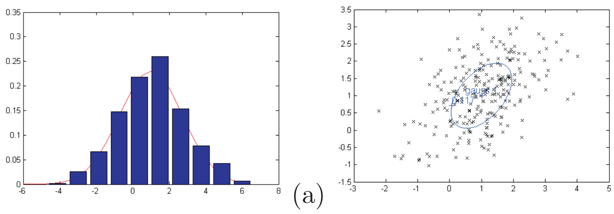
Statistical Pattern Recognition

- Sampling a Gaussian Distribution
  1. Univariate case
  2. Bivariate case
- Maximum Likelihood Estimation
- Bayes Classifier

Sampling a Gaussian Distribution

1. A data set with only one feature **x** (univariate distribution) has the following Gaussian parameters: mean 1 and variance 3. Construct a model using the function `struct` with the two parameters. Plot the probability distribution function (pdf) of the Gaussian in the range $[-6 : 0.5 : 6]$ using the function `pdfgauss`. Generate 500 samples of the model (e.g gsamp(model,500). Plot the histogram.

2. Repeat previous exercise for a data set with two features x and y(bivariate distribution) with Mean = [1;1]; and Cov = [1 0.5; 0.5 1]. Use function pgauss which vizualizes a set of bivariate Gaussians.



(a) (b)

Figura 1: Gaussian Distribution(a)Univariate case (b) Bi-variate

3. Consider the Ripley data set contained in a data directory of your STPRTool Pattern Recognition tool software. The data is in the Matlab file 'riply_trn.mat' and contains labelled points corresponding to two classes. It can be loaded into your workspace by typing:
`data = load('riply_trn'); % load labeled (complete) data`
Estimate the Maximum Likelihood (ML) of a Mixture Gaussian Model. Use function `mlcgmm` to construct the model and functions `pgauss` and `pgmm` to visualize the model. With these functions the graphs in Figure 2 are obtained.

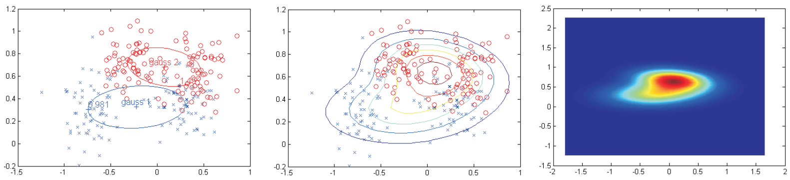


Figura 2: ML Estimation on Ripley Data

4. With the Ripley data set from previous exercise construct a Bayesian classifier. Test the classifier with the Ripley data test set:
`data = load('riply_tst'); % load labeled (complete) data`
Compute decision boundary with:
Function `bayesdf`

   Synopsis: `gauss_model = mlcgmm(trn)`
   `quad_model = bayesdf(gauss_model)`

   Description:
   This function computes parameters of decision boundary of the Bayesian classifier with the following assumptions:

   - 1/0 loss function (risk = expectation of misclassification)
   - Binary classification
   - Class conditional probabilities are multivariate Gaussians.

5. Repeat previous exercise on the same data with Bayesian Classifier function given by `bayescls`. This function implements the classifier minimizing the Bayesian risk with 0/1-loss function. It corresponds to the minimization of probability of misclassification. The input vectors X are classified into classes with the highest a posterior probabilities computed from given model.

6. Consider the first two classes of the `Cork Stoppers.xls` data set described by the features `ART` and PRT.

   (a) Compute the decision boundary
   (b) Compute the Bayes error
   (c) Consider the three most discriminative features. Compute the Bayes error for two and three classes. Compare results.
   (d) Consider the three least discriminative features. Compute the Bayes error for two and three classes. Compare results.

7. Consider the `Fruits` images data set. Process the images in order to obtain the features (a picture program, such as Corel Draw can be used for this purpose) - . Design a Bayesian Classifier for the 3-class fruit discrimination. Comment the results obtained.

Figura 3: Three Fruits in your dataset

8. A physician would like to have a very simple rule available for screening out the carcinoma situations from all the other situations, using the same diagnostic means and measurements as in the `Breast Tissue` data set.

    a) Read your data from excel file `breast_tissue.xls`.
    You will need to convert a class cell array to numeric value. Define a struct `CLASS (Name_Class, Num_Class)`. You can the use the following MatLab Code (or a similar one):

    ```
     dim=size(BT_CLASS,2);
    CLASS.Name_Class = BT_CLASS;
      for i=1:dim
       if strcmp(BT_CLASS{i}, 'car' )
        CLASS.Num_Class{i}=1
       elseif strcmp(BT_CLASS{i},'fad')
        CLASS.Num_Class{i}=2
       ...
        else
       strcmp(BT_CLASS{i}, 'adi')
        CLASS.Num_Class{i}=6
       end
      end
      CLASS.dim=dim;
    ```

    b) First define a variable `BT_C` with class values `1` and `2` corresponding to carcinoma and all other cases respectively.

    c) Build Train (trn) and Test (tst) data sets.
    Hint: Use `randperm` function for generating random permutations of data set. Be careful with MatLab set up dimensions when using `randperm`.

    d) Using the `Breast Tissue`, find a Bayesian classifier with the three most discriminant and three less discriminant features for the carcinoma classification versus all other cases. Hint: Use Kruskall-Wallis method. Sort ranks and choose features.

    e) Obtain training set and test set error estimates. Hint: Use `cerror` on each of the training and test data sets (`trn` and `tst` respectively).

    f) Design a Bayesian classifier with all features and compare the results

Técnicas de Reconhecimento de Padrões
Pattern Recognition Techniques

2009/2010

Practical Exercises : *Class # 4*

Data Sets from:
"Pattern Recognition: Concepts, Methods and Applications"

Topics

Statistical Pattern Recognition

- Non-Parametric Methods
  1. K-Nearest Neighbour (K-NN)
  2. ROC Curves
- Use SPRTool Matlab Functions `knnrule`,`knnclass`,`roc`

K-Nearest Neighbour

1. Perform a K-NN classification of the `Cork Stoppers Data Set` in order to discriminate between classes ω_1, ω_2 and ω_3 .

2. Perform a K-NN classification of the `Breast Tissue Data Set` in order to discriminate carcinoma from all the other cases.

3. Perform a K-NN classification of the `Iris Data Set` in order to discriminate the three classes of IRIS (setosa, versilcolor, virginica).

Iris_setosa	Iris_versicolor	Iris_virginica
		

Figura 1: Iris Data Set

4. Perform the k-NN classification of the Rocks Data,using two classes: *granites,diorits, schists* vs. *limestones, marbles, breccias*.

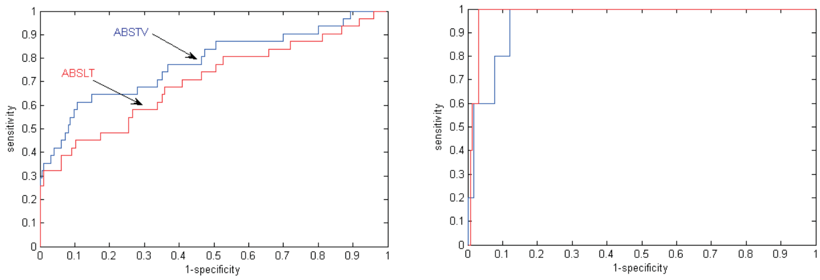


Figura 2: ROC Curves: `FHR Apgar data set`

(a) give an estimate of the k neighbours that should be used

(b) for the previously estimated k what is the expected deviation of the asymptotic error for the K-NN classifier from the Bayes error.

ROC Curves

1. Exercise on ROC Curves

   (a) Explain why all ROC Curves start at (0,0) and finish at (1,1) by analysing what kind of situations they correspond to?

   (b) The Excel file `SignalNoise.xls` contains 100 samples of signal + noise. The arrival times of the signal impulses have a Poisson distribution. Change the value of the detection threshold and observe the changes performed on the detections. Plot the the computed sensibility and specificity using 10 thresholds.

   (c) Consider the `Breast Tissue Data Set`. Use the ROC curve approach to determine single features that discriminate carcinoma cases from all other cases.

   (d) Calculate the ROC curves for the indexes 1 and 5 for the `FHR Apgar data set`, using combinations of features.

Técnicas de Reconhecimento de Padrões
Pattern Recognition Techniques

2009/2010

Practical Exercises : *Class # 5*

Data Sets from:
"Pattern Recognition: Concepts, Methods and Applications"

Topics

Hierarchical Tree Clustering

K-Means Clustering

Cluster Validation

Hierarchical Tree Clustering Algorithms

Hierarchical Tree Clustering Algorithms use linkage rules to produce hierarchical sequence of clustering solutions.

1. Use Matlab files `pdist.m`, `linkage.m` and `dendrogram.m`. Type Help to know more about these functions.A summary of main important points are specified below.

```
Y = PDIST(X, DISTANCE) computes Y using DISTANCE.  Choices are:

        'euclidean'   - Euclidean distance
        'seuclidean'  - Standardized Euclidean distance, each coordinate
                        in the sum of squares is inverse weighted by the
                        sample variance of that coordinate
        'cityblock'   - City Block distance
        'mahalanobis' - Mahalanobis distance
        'minkowski'   - Minkowski distance with exponent 2
        'chebychev'   - Chebychev distance (maximum coordinate difference)

 LINKAGE Create hierarchical cluster tree.

     Z = LINKAGE(Y, method) creates a hierarchical cluster tree using
    the specified algorithm. The available methods are:

       'single'    --- nearest distance
```

```
'complete'  --- furthest distance
'average'   --- unweighted average distance (UPGMA) (also known as
                group average)
'weighted'  --- weighted average distance (WPGMA)
'centroid'  --- unweighted center of mass distance (UPGMC) (*)
'median'    --- weighted center of mass distance (WPGMC) (*)
'ward'      --- inner squared distance (minimum variance algorithm)
```

2. Consider the `cluster.xls` data set. Use `function scatter.m` to draw the scatter plot of:

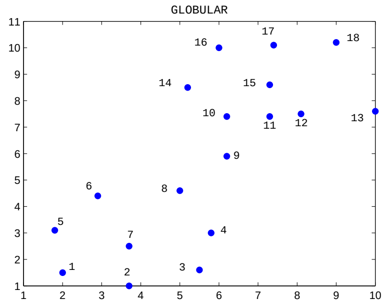
   (a) globular

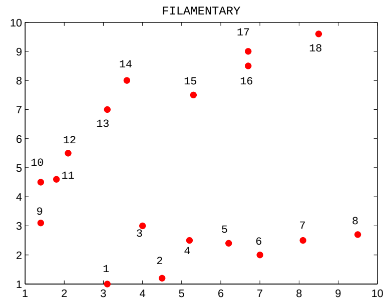
   (b) filamentary

   (c) +Cross data

   (d) xCross data

3. In the scatter diagram above for `Globular` and `Filamentary` give a number to each point in order to better understand the tree clustering solution. Example is as follows:

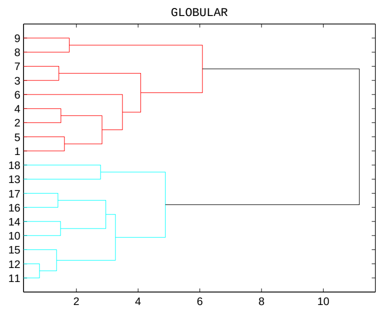


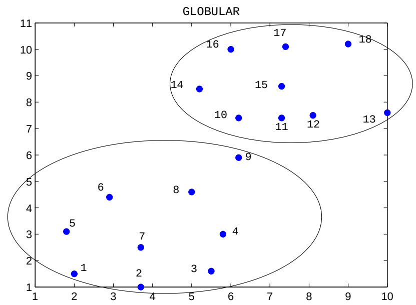


4. Use `pdist='euclidean'` and `linkage='complete'` on `Globular`. Run the following Matlab code:

```
Y=pdist(GLOB_DATA,'euclidean');
Z=linkage(Y,'complete');
[H,T]=dendrogram(Z,'colorthreshold','default','orientation','right');
```

Here is a possible solution where the Globular Clusters are clearly identified.





5. Repeat the exercise to obtain more filamentary clusters in the `Filamentary` data. Adjust the distances and rule of merging (linkage). Hint: Use `euclidean` distance and `single` linkage rule.

6. Determine the tree clustering solutions of the `+Cross` using the WPGMA linkage rule with the euclidian, city block and Chebychev norms. Explain results.

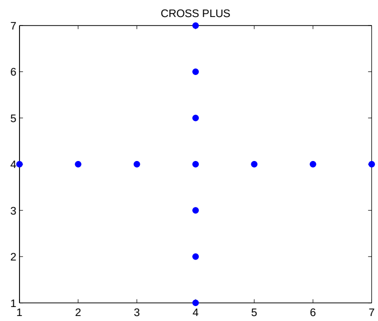


Figura 1: +Cross data

7. Determine the tree clustering solutions of `xCross` data using the WPGMA linkage rule with the city-block and Chebychev norms. Explain results and compare them with those obtained in previous exercises.

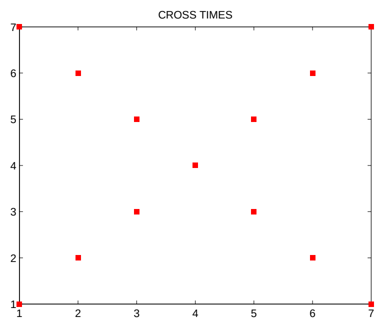


Figura 2: xCross data

8. Determine the tree clustering solutions of `Globular` data using UPGMA and WPGMA linkage rules with the Euclidian norm. Explain results and compare them with those obtained in previous exercises.

9. Determine the tree clustering solutions of `Filamentary` using the ward method with the city-block norm. Explain results and compare them with those obtained in previous exercises.

10. Find the following tree clustering solution for the `Crimes` data set using complete linkage rule.

```
Cluster1={Lisboa, Faro, Leiria, Guarda, Vila Real, Évora,
          Portalegre, Castelo Branco}
Average incidence crimes  against property and people

Cluster2={Viana do Castelo, Setubal, Aveiro}
High incidence crimes against property;above avg against people

Cluster3={Coimbra, Bragança, Santarém, Braga, Beja,Viseu,Porto}
High incidence crimes against property;below avg against people
```

11. Find a four cluster solution for the `Food` data set. See PR book Figure 3.16.

K-means Clustering

K-means clustering) is a Centroid Adjustment Algorithms whose main objective is to adjust prototypes centroids describing the clusters.

1. Use the `kmeans.m` available in the STPRTool to run the algorithm in Riply Data.

```
 data = load('riply_trn');
   [model,data.y] = kmeans( data.X, 4 );
   figure; ppatterns(data);
   ppatterns(model.X,'sk',12); pboundary( model );
```

2. You should be able to visualize the picture below which marks the cluster centers.

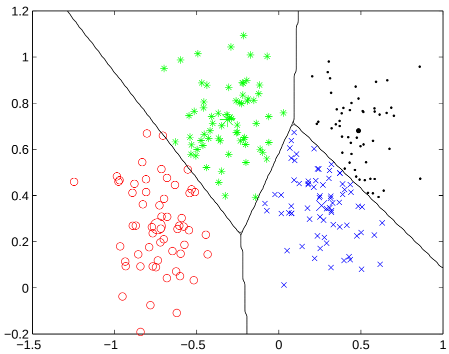


Figura 3: Kmeans Clustering on Riply Data

3. Build the training and tests sets as well as the respective models using `kmeans.m` for the following Data Sets:

(a) Data Sets

CORK_STOPPERS.xls
BREAST_TISSUE.xls

(b) Visualize the solutions for both.

4. Consider `Rocks` data set. Find the three clusters identified in Figure below. Use two Principal Components and K-means (c=3 Clusters)

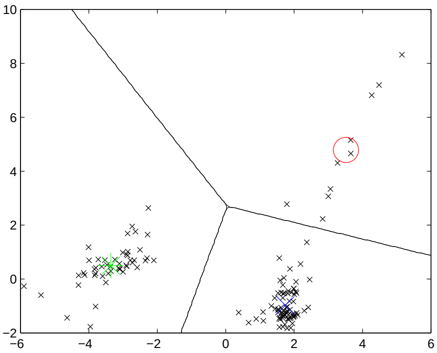


Figura 4: Three cluster solution (c=3) in the Rocks data set

Técnicas de Reconhecimento de Padrões
Pattern Recognition Techniques

2009/2010

Practical Exercises : *Class # 6*

Data Sets from:
"Pattern Recognition: Concepts, Methods and Applications"

Topics

Kernel Learning

- Support Vector Machines
  1. hard margin
  2. soft margin
- Use SPRTool Matlab Functions `smo, svmquadprog, svmclass, oasvm,`

Support Vector Machines

Figura 1: Cork Stoppers Image Defects

1. Apply SVM to the `Cork Stoppers Data Set` (Figure 1) in order to discriminate between classes ω_1, ω_2. Make experiments with `rbf` kernel with varying parameters (`0.1, 1, 10, 100`). Likewise use several values of constant C (e.g, 1, 10, 100). Determine the number of SVs (Support Vectors) for the best combination of parameters. Use optimization function `smo` and svm classifier `svmclass`. Give the error in the design (training) data set Pe_d. Plot the ROC curve (Figure 2).

2. Perform the SVM classification for discriminating the three classes ω_1,ω_2 and ω_3 of above problem. Use multiclass svm function `oaosvm, ''one against all´´`. In setting options use also parameter 'verb' (verbosity) to 1:
`options=struct('ker','rbf','arg',100,'C',100,'verb', 1)`
Evaluate the prediction error.

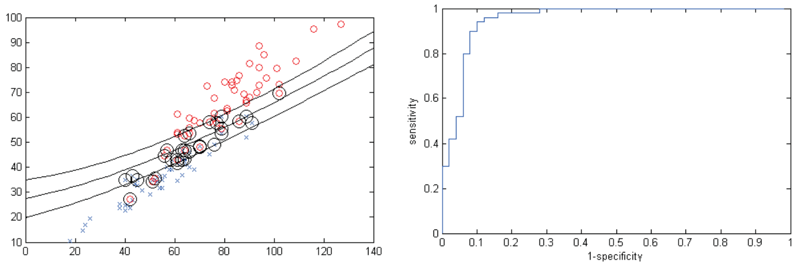


Figura 2: Cork Stoppers Classification

3. Design a SVM for classification of the Rocks data into two classes: `granite vs limestone+marbles`. Use features SiO2, CaO and determine experimentally the kernel with best generalisation (`linear, rbf, or polynomial`).

4. Perform SVM classification to the `Breast Tissue Data Set` in order to discriminate carcinoma from all the other cases. Determine experimentally the best features, best kernel, optimal parameter C, number of SVs and prediction error as in previous examples.

5. Perform a SVM classification of the `Iris Data Set` in order to discriminate the three classes of IRIS (Setosa, Versilcolor, Virginica). Determine experimentally the best kernel, optimal parameter C, number of SVs and prediction error as in previous examples.

6. Perform a SVM classification of the `CTG Cardiographic Data` which contains 2126 measurements and classifications of foetal heart rate (FHR) signals in order to discriminate the three classes (Normal, Suspect, Pathologic). Use a reduced data set. Determine experimentally the best features, best kernel, optimal parameter C, number of SVs and prediction error as in previous examples.