Set of Questions

Exploração de Dados

Ana Maria Tomé
&
Raquel Sebastião

Important: this document contains a set of questions covering the topics addressed in the course "Data Mining". The slides of the lectures have the main concepts. Much more details can be found in the suggested bibliography.

(Versão de December 31, 2022)

Contents

1 Data sets and Features

The following questions are addressing

- the basic terminology about features/atributes, objects/examples and data sets;
- the univariate and bivariate analysis;
- the proximity measures (to calculate similarity or dissimilarity between objects);
- the data transformations;

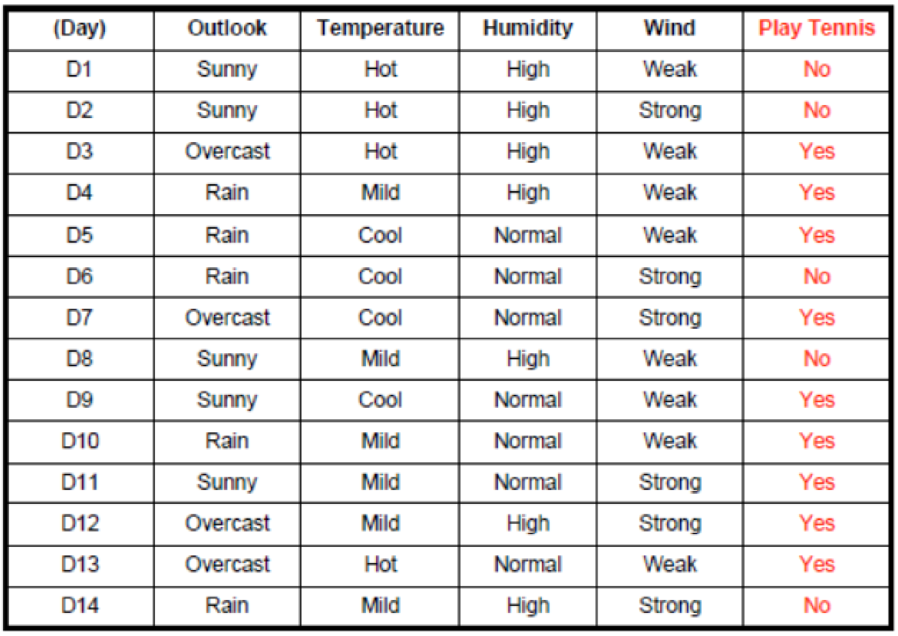
Problem 1.1
Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative. If you consider that some cases may have more than one interpretation, briefly indicate your explanation. Example: Age in years. Answer: Discrete, quantitative.

1. Time in terms of AM or PM.
2. Brightness as measured by a light meter.
3. Brightness as measured by people's judgments.
4. Angles as measured in degrees between 0° and 360° .
5. Bronze, Silver, and Gold medals as awarded at the Olympics.
6. Number of patients in a hospital.
7. Ability to pass light in terms of the following values: opaque, translucent, transparent.
8. Military rank.

Problem 1.2
The figure 1.1 shows examples of a very popular data set in machine learning [2]

1. How many instances (or examples) has the table?
2. How many features (or attributes) has the data set?
3. Consider that the feature ”play” is the label (class) to be predicted. What is the dimension of the problem?
4. The first column is not used on data analysis tasks. Why?

5. All the features (or attributes) are categorical. Identify the nominal attributes and the ordinal attributes.
6. Convert the categorical attribute to numeric values. Explain the procedure.



(Day)	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Figure 1.1: Data set

Problem 1.3
Consider the following data set

$$\mathbf{D} = \left(\begin{array}{c|cc} & X & Y \\ \hline \mathbf{a}_1^T & 2 & 0.8 \\ \mathbf{a}_2^T & 5 & 2.4 \\ \mathbf{a}_3^T & 8 & 5.5 \end{array}\right)$$

1. Represent the data in the feature space.
2. Calculate mean vector $\boldsymbol{\mu}$. Sol.: $\boldsymbol{\mu} = [5 \quad 2.9]$.
3. Center the data, i.e, all features should be zero mean. Sol.: $\mathbf{Z} = \begin{pmatrix} -3 & -2.1 \\ 0 & -0.5 \\ 3 & 2.6 \end{pmatrix}$
4. Show that the data centering can be a matrix manipulation

$$\mathbf{Z} = \mathbf{D} - \mathbf{1}\boldsymbol{\mu}^T$$

where $\mathbf{1}$ is a vector of ones.

5. Calculate covariance matrix. What is the meaning of each entry of the covariance matrix? Sol.: $\mathbf{C} = \begin{pmatrix} 9 & 7.05 \\ 7.05 & 5.71 \end{pmatrix}$

6. Calculate the correlation coeficcient (Pearson's correlation) between the feature X and the feature Y. Sol.: $\rho_{\mathbf{XY}} = 0.98$

Problem 1.4

Consider the data set of exercise 1.3 and calculate

1. The following proximity (similarity and dissimilarity) measures
   - the Euclidean distance (L_2 norm) between $\mathbf{a}_1$ and $\mathbf{a}_2$. Sol.: $d_E(\mathbf{a}_1, \mathbf{a}_2) = 3.4$
   - the Manhathan distance (L_1 norm) $\mathbf{a}_1$ and $\mathbf{a}_2$. Sol.: $d_M(\mathbf{a}_1, \mathbf{a}_2) = 4.6$
   - the cossine distance $\mathbf{a}_1$ and $\mathbf{a}_2$. Sol.: $d_C(\mathbf{a}_1, \mathbf{a}_2) = 0.9978$

2. Two new examples are inserted in the data set. These new points are defined as

$$\mathbf{a}_4 = 2\mathbf{a}_2 \qquad\qquad \mathbf{a}_5 = \mathbf{a}_2 + 2\mathbf{1}$$

   Repeat 1 for the pairs $(\mathbf{a}_1, \mathbf{a}_4)$ and $(\mathbf{a}_1, \mathbf{a}_5)$. Compare with the results obtained for the pair $(\mathbf{a}_1, \mathbf{a}_2)$. Justify the results. Par. Sol.: $d_E(\mathbf{a}_1, \mathbf{a}_4) = 8.9$, $d_M(\mathbf{a}_1, \mathbf{a}_4) = 12$, $d_C(\mathbf{a}_1, \mathbf{a}_4) = 0.9978$ $d_E(\mathbf{a}_1, \mathbf{a}_5) = 6.2$, $d_M(\mathbf{a}_1, \mathbf{a}_5) = 8.6$, $d_C(\mathbf{a}_1, \mathbf{a}_5) = 0.9837$

Problem 1.5

Figure 1.2 represents $2D$ data set. The first plot (left) represents the original values while the two plots represent the data after applying to each feature a transformation.

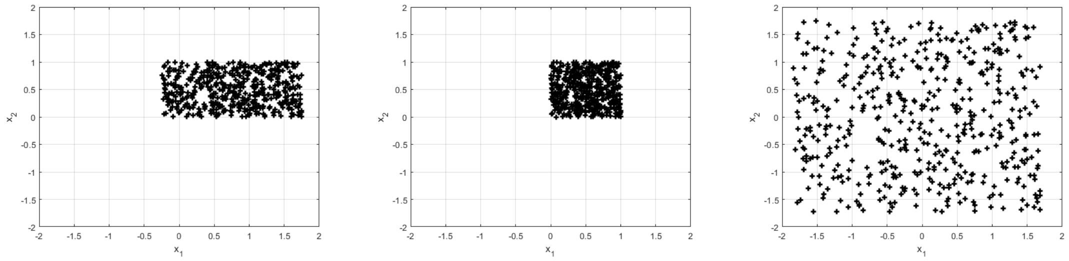


Figure 1.2: Data set with 2 features:Left: Original data; Midle and Right: after normalization

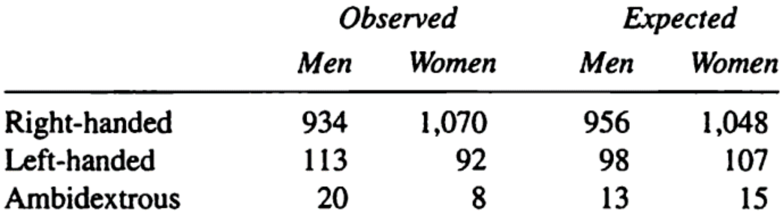
1. By visual inspection give an approximation to the mean values of the features in each plot. Geometrically localize the mean vector of the data set in the figures.

2. The normalization strategies were the zscore and min-max.Describe the normalization procedures.

3. By visual inspection can you identify the plot corresponding to min-max normalization? Justify your answer.

4. Which of the characteristics of the remaining plot indicates that zscore was applied.

Problem 1.6
The following contingency table describes how two groups Men and Women are distributed according to the feature Handness.



	Observed		Expected	
	Men	Women	Men	Women
Right-handed	934	1,070	956	1,048
Left-handed	113	92	98	107
Ambidextrous	20	8	13	15

Figure 1.3: Contingency tables between Gender and Handness. Left: Observed in a group of 2237 persons; Right: Expected (note that the values are rounded to integer)

1. What is the type of the feature Handness?

2. Calculate the values of table on the right. The values are calculated assuming independence between Handness and Gender. Note that if A and B are independent then $P(AB) = P(A)P(B)$

3. Calculate the chi-square value (χ^2).

2 Dimension Reduction Techniques

The following questions are related with principal component analysis and linear discriminant analysis.

The main goal of these methods is to find vector basis to project the data. The projection operation can be written as

$$\mathbf{p} = \mathbf{U}_L^T \mathbf{x}$$

The L columns of the matrix $\mathbf{U}$ are the vector basis. The feature vector $\mathbf{x}$ with dimension D is projected into the basis system. The $l - th$ entry of the vector $\mathbf{p}$ represents the projection into $l - th$ column of $\mathbf{U}_L$. If $L << D$ a dimension reduction is achieved.

Problem 2.1
Condider the following vectors

$$\mathbf{a} = (1/\sqrt{2} \quad 1/\sqrt{2})^T \quad \mathbf{b} = (-1 \quad -1)^T$$

1. Represent the two vectors in 2D space. Calculate the angle between the two vectors. Par. Sol.: $\theta = \pi$
2. Calculate the length of the vectors and the angle with positive $x - axis$. Sol.: $(||\mathbf{a}|| = 1$, $(\theta_a = \pi/4$, $(||\mathbf{b}|| = \sqrt{2}$, $(\theta_b = 5\pi/4$
3. Geometrically calculate the angle between the two vectors.
4. Calculate the dot product $\mathbf{b}^T\mathbf{a}$. Sol.: $\mathbf{b}^T\mathbf{a} = -\sqrt{2}$
5. Geometrically explain the meaning of the dot product.
6. The dot product can be zero. Calculate $\mathbf{c}$ such that $\mathbf{c}^T\mathbf{a} = 0$. Explain geometrically the result of this operation. Par. Sol.: $\mathbf{c} = (-1 \quad 1)^T$ (for example)

Problem 2.2
The figure 2.1 shows examples of a data set with two features (2D dataset) and the main directions of the data. The eigenvectors are the columns of the following matrix

$$\mathbf{U} = \begin{pmatrix} -0.44 & -0.88 \\ -0.88 & 0.44 \end{pmatrix}$$

the corresponding eigenvalues are 1.7 e 0.009, respectively.

1. Represent the eigenvectors on the feature space.

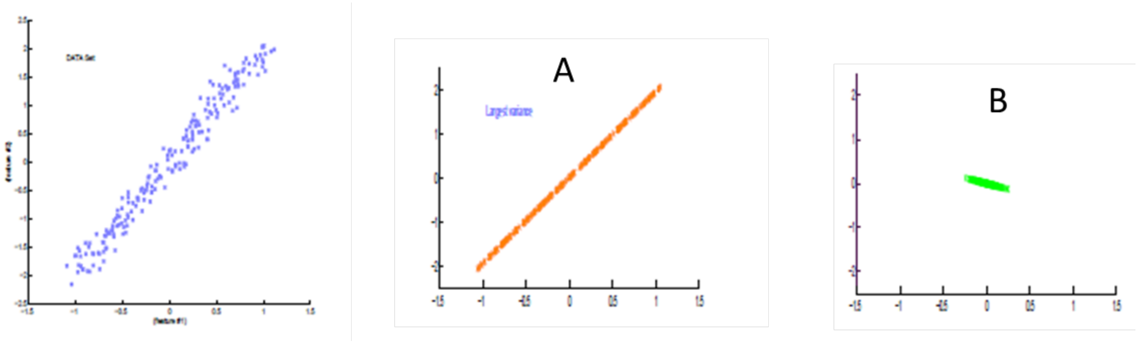


Figure 2.1: Data set and principal directions

2. Consider the following $\mathbf{x}_r = \mathbf{u}_1\mathbf{u}_1^T\mathbf{x}$, $\mathbf{u}_1$ is the first column of $\mathbf{U}$ and $\mathbf{x}$ is one of the points of the data. If the operation is repeated for all data points one of the plot (middle or right) is obtained. Which one?

Problem 2.3
The figure 2.2 shows examples of a data set with two features (2D dataset). The data set is centered.

1. Explain the statement: ”The data set is centered”.

2. Applying the SVD algorithm to the 1000×2 data matrix $\mathbf{Z}$ the right eigenvector matrix is

$$\mathbf{U} = \begin{pmatrix} 0.046 & 0.999 \\ 0.999 & -0.046 \end{pmatrix}$$

corresponding to the following non-zero singular values 33.7 and 17.9 respectively. Represent the eigenvectors on the feature space.

3. Consider that you project the data into the eigenvector related with the largest singular value. The values of the projections of the two clouds of points will have different ranges? Justify.

Problem 2.4
The toy data set of figure 2.2 is now represented using also the label information of the elements (see the figure 2.3 on the left). The class red is centered on $(0.11, -0.43)^T$ and

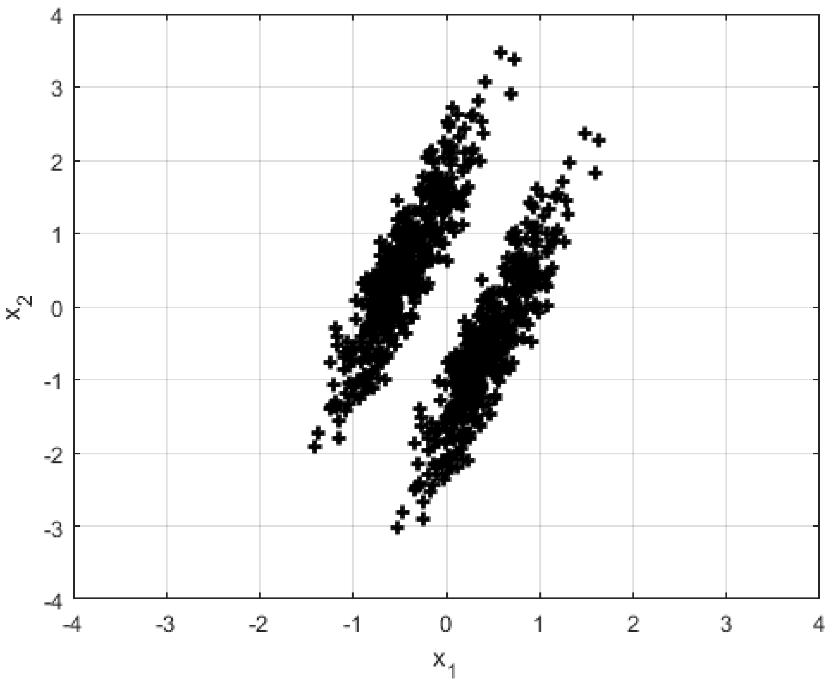


Figure 2.2: Data set: feature vector $\mathbf{x} = (x_1 \quad x_2)^T$

the class black is centered on $(-0.79, 0.65)^T$. The data set was used to calculate the Fisher Discriminant model that is described by

$$\mathbf{w} = \begin{pmatrix} 0.05 & -0.02 \end{pmatrix}^T$$

All the examples of the data set were projected into $\mathbf{w}$ and into the first column of the matrix $\mathbf{U}$ (see 2.3).

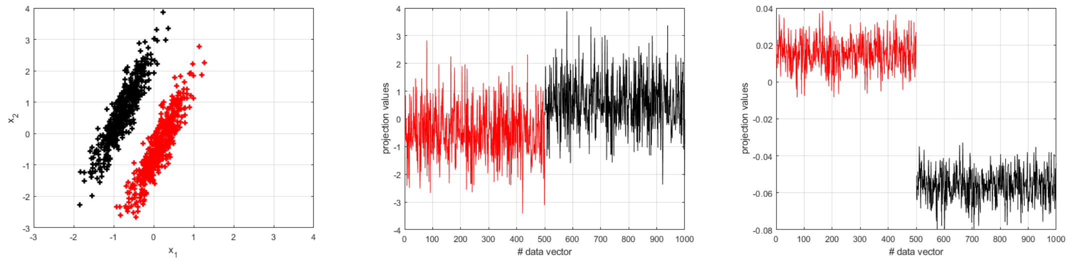


Figure 2.3: Data set: feature vector $\mathbf{x} = (x_1 \quad x_2)^T$. left: Two classes: red and black, respectively; middle: SVD projections; right: Fisher projections.

1. Consider the projection values into the two models and comment the following issues.
   - The SVD projection values do not allow to discriminate the two classes.
   - The Fisher projection values can be used to separate the two classes. How?
2. With the Fisher discriminant is possible to construct a decision rule which indicates the class of an example $\mathbf{x}$. The sign of the following calculation
   $$\mathbf{w}^T\mathbf{x} + b$$
   can be related with the classes. Calculate b, without using the graphic output, and explain the rule.
3. Comment: "Fisher Discriminant can be considered a pre-processing model to reduce the dimension of the feature vectors and a classification model".

Problem 2.5

The figure 2.4 illustrates data set where the examples belong to two classes. The data set was used to calculate: PCA, LDA e KPCA models.

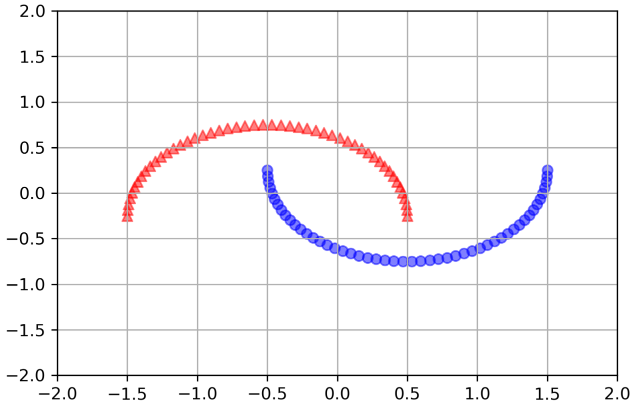


Figure 2.4: Data set with two classes (blue and red)

The PCA is described by the eigenvector matrix

$$\mathbf{U} = \begin{pmatrix} -0.95 & 0.3 \\ 0.3 & 0.95 \end{pmatrix}$$

and the corresponding eigenvalues 3.0 and 2.1, respectively. The LDA model is described by the vector

$$\mathbf{w} = \begin{pmatrix} 0.26 \\ -0.97 \end{pmatrix}$$

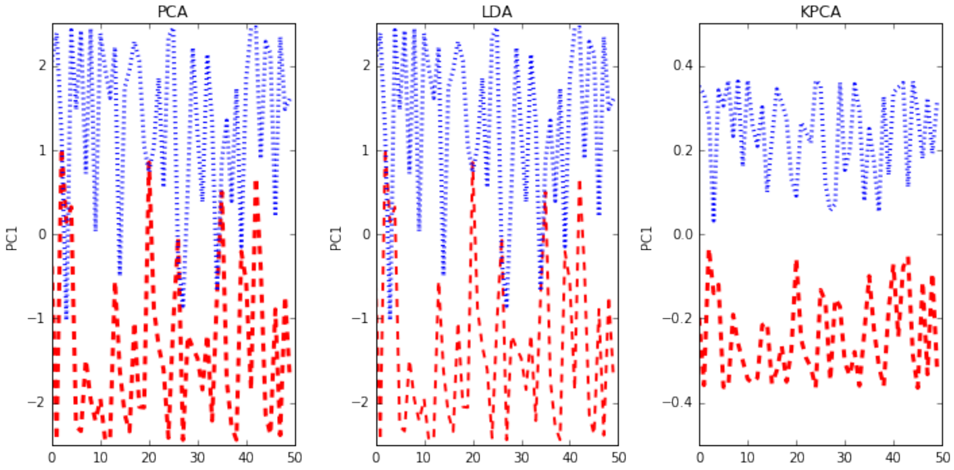


Figure 2.5: Projections of the examples of the data set (see figure 2.4) into the directions of PCA, LDA and KPCA. The values of each class are represented by distinct colours(and symbols) . The x-axis is the index of each example (within each class).

1. Only one of the models is computed taking into account the class information.
2. Why LDA model has only one vector?
3. The first eigenvector of PCA is related with the largest eigenvalue. What means the direction of this vector in data space?
4. Explain the relation between SVD and PCA.
5. Compute the projection of $[-1, 1]^T$ in both LDA and PCA.
6. The figure 2.5 illustrate the values of the projections of the three models.
    - Comment the outcomes using the properties of the models.
    - With the projections is possible to design the following decision rule

$$\begin{cases} p > \alpha & class = 1 \\ p < \alpha & class = -1 \end{cases}$$

    Where p is the projection value into one of the models. What is value of α ? The decision is correct for all examples? Which of the projection models leads to correct decisions for all examples?

3 Clustering Algorithms

Clustering Algorithms are unsupervised machine learning techniques. They are based on different strategies and the most popular algorithms are

- K-means: a proximity measure should be chosen (Euclidean distance is the most popular)
- Hierarchical Algorithms: a proximity measure between sets (single linkage , average linkage and so on).
- Density Based Clustering.

Problem 3.1
The figure 3.1 represents examples of a data set. The K-means algorithm is applied to form $K = 2$ clusters

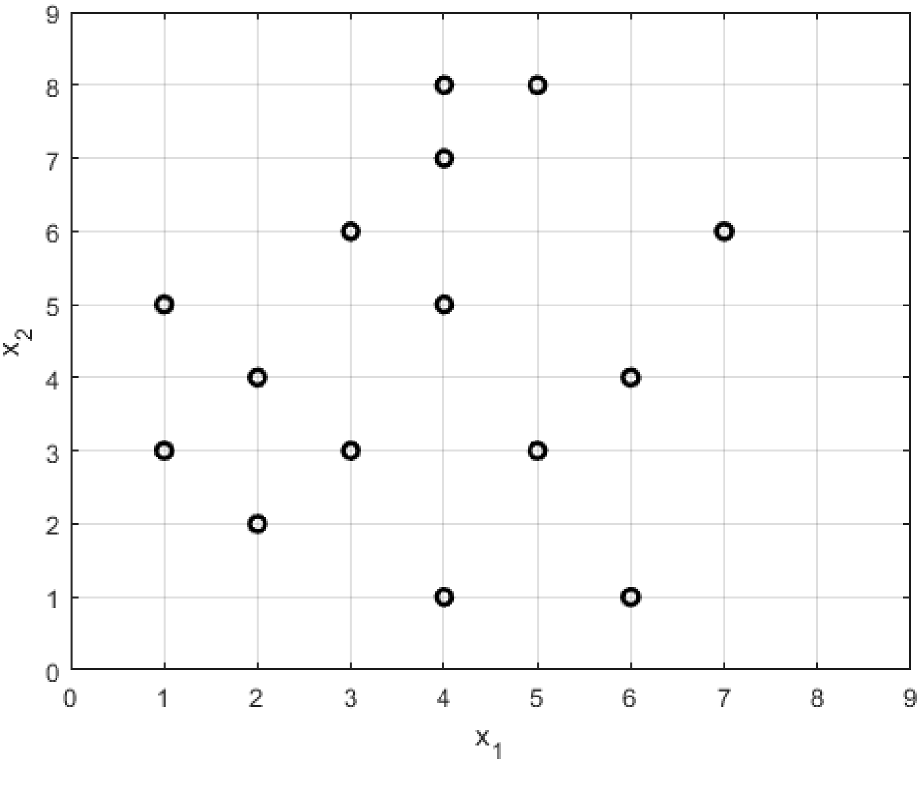


Figure 3.1: Similarity measures

1. What are the main steps of K-means?

2. If the initial values of the centroids are: $\mathbf{c}_1 = (3,2)^T$ and $\mathbf{c}_2 = (6,7)^T$,
   - identify the examples of each cluster using the Euclidean distance
   - update the value of the centroid. Sol: $(3, 2.75)^T$ and $(4.7, 6.3)^T$
3. With the new values for the centroids the two clusters have the same examples?
4. Define a criterium to stop the algorithm.
5. Starting the algorithm with centroids: $\mathbf{c}_1 = (1,5)^T$ and $\mathbf{c}_2 = (7,6)^T$ is it expected to obtain the same clusters?

Problem 3.2
K-means is often used to segment colored images.

1. Considering that the image is RGB, explain how to form the feature vector. What is the dimension of the feature vector?
2. Assuming that the image is outdoor image with a very large area with green grass. If you want to identify the grass area in the image how do apply K-means?

Problem 3.3
Hierarchical Clustering algoritms make use of proximity measures between sets. Define the similarity measures illustraded on the figure 3.2.

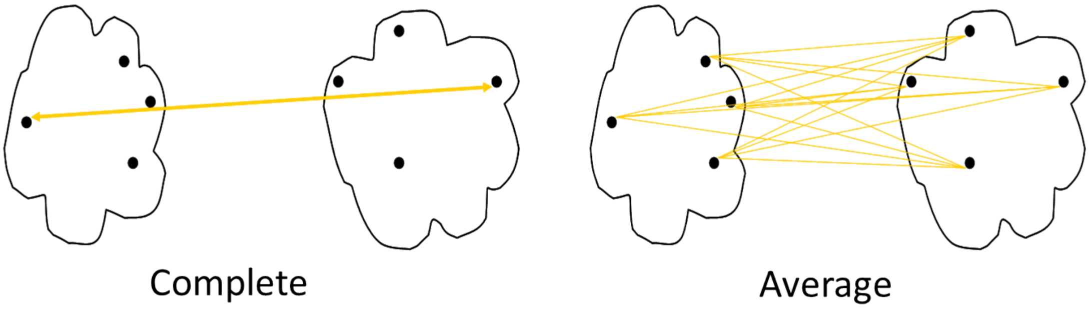
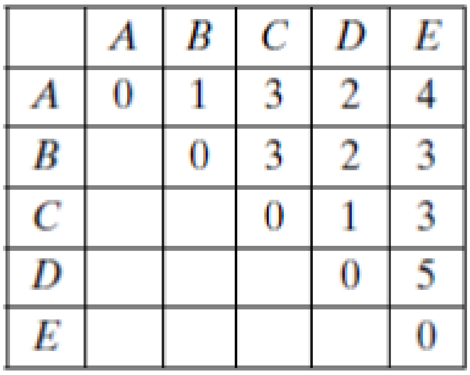


Figure 3.2: Similarity measures

Problem 3.4
Use the distance matrix represented in the figure 3.3 to perform single and complete linkage hierarchical clustering (agglomerative). The dendrogram should clearly show the order in which the objects are merged.



	A	B	C	D	E
A	0	1	3	2	4
B		0	3	2	3
C			0	1	3
D				0	5
E					0

Figure 3.3: Distance matrix

Problem 3.5
Consider the following 1D data set

$$X = \{2, 4, 5, 9, 10\}$$

Using the Euclidean distance, perform single and complete linkage hierarchical clustering (agglomerative). The dendrogram should clearly show the order in which the objects are merged.

Problem 3.6
The figure 3.4 represents the application of two clustering algorithms. Which of them can be the DBSCAN. Explain the basic principles of the algorithm.

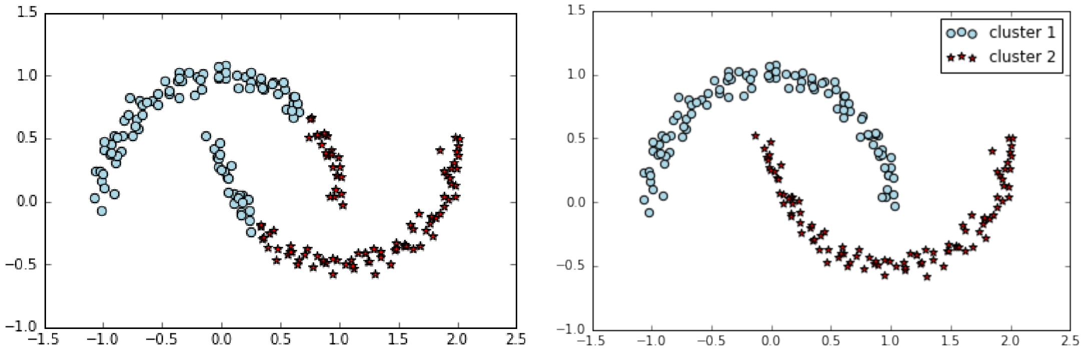


Figure 3.4: Data clustered by two different algorithms

4 Classification

Classifications strategies based on the calculation of the decision surface. Linear (hyperplanes) and Non-linear decision surfaces are ilustrated with different algorithms.

4.1 Linear Classifiers

The following questions are related with the calculation of models of classification explained with linear discriminant functions. The binary classification problem is solved by the following function

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

The parameters of the model are $\mathbf{w}$ and b (also called w_0). Given a feature vector $\mathbf{x}$, the sign of the function $g(\mathbf{x})$ indicates the class of the feature vector. The following exercises address the following methods: Fisher discriminant, perceptron, Least Squares and SVM.

Problem 4.1

Consider the data set of the table 1

ID	X_1	X_2	Label (d)
1	4	2.8	1
2	3.5	4	1
3	2.5	1.1	-1
4	2	2.1	-1

Table 1: Data set with 4 examples and dimension 2

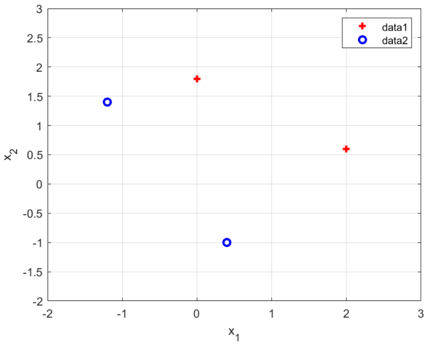
1. Calculate the mean of each class and the global mean. Sol: $(2.25,\ 1.6)^T; (3.75,\ 3.4)^T; (3,\ 2.5)^T$
2. Calculate the within Class scatter matrix. Sol: $\begin{pmatrix} 0.25 & -0.55 \\ -0.55 & 1.22 \end{pmatrix}$
3. Calculate the between class scatter matrices. Sol:$\begin{pmatrix} 2.25 & 2.7 \\ 2.7 & 3.24 \end{pmatrix}$
4. Calculate the vector $\mathbf{w}$ [1].
   - Using an eigedecomposition. Sol:$(0.91,\ 0.41)^T$. Note use Python or any other tool to calculate the eigendecomposition.
   - Using the 2 class simplification. $(1128,\ 510)^T$

[1] The inverse of a 2 × 2 matrix http://www.mathcentre.ac.uk/resources/uploaded/sigma-matrices7-2009-1.pdf

5. What is the difference between the vectors $\mathbf{w}$.
6. Calculate b. Sol: -4659 or -3.76.

Problem 4.2
The perceptron algorithm can be used to learn the binary classification problem illustrated on the following figure. The data is also written in the table.

ID	X_1	X_2	Label (d)
1	-1.2	1.4	-1
2	0.4	-1	-1
3	0	1.8	1
4	2	0.6	1



Assume that the decision hyperplane passes by the origin of the feature space and the initial weight vector is $\mathbf{w} = (1,\, 1)^T$

1. Represent the hyperplane in the feature space. Suggestion: calculate the angle of the hyperplane with $x - axis$. Sol.: $\theta = 3\pi/4$
2. The first element of the data (see table) is correctly assigned to class -1?
3. Apply the perceptron updating rule to calculate the new weight vector $\mathbf{w}$. Sol.: $\mathbf{w} = (1.8,\, 0.6)^T$
4. Write a short program to find out how many iterations are needed for the model to learn the classes of all elements of the data set. Par. Sol.: 3 iterations

Problem 4.3
The data set of the table 2 was used to train a linear SVM. The Lagrangian values are the outputs of the training algorithm

1. Comment the statement: ”the data set is linearly separable”.
2. Explain: ”The elements of the data set whose Lagrangian have non-zero values are defining two hyperplanes paralell to the decision hyperplane”.
3. Calculate $\mathbf{w}$ defining the direction of the decision hyperplane. Sol: $\mathbf{w} = (0.846,\, 0.3852)^T$.
4. Calculate the value of b which defines the decision rule. Note that at the margins the $\mathbf{w}^T\mathbf{x} = \pm 1$. Sol: $b \approx -3.5$.
5. What is the class of $\mathbf{z} = (3,\, 3)^T$?

ID	X_1	X_2	d(label)	Lagrangian(λ)
1	4	2.9	1	0.414
2	4	4	1	0
3	1	2.5	-1	0
4	2.5	1	-1	0.018
5	4.9	4.5	1	0
6	1.9	1.9	-1	0
7	3.5	4	1	0.018
8	0.5	1.5	-1	0
9	2.0	2.1	-1	0.414
10	4.5	2.5	1	0

Table 2: Data Set and Lagrangian values after training a linear SVM

4.2 Non-Linear Classifiers

Non-linear decision surfaces will be learned by different approaches. The SVM classifier with radial and polynomial kernel functions, multilayer feedforward neural networks and random forrest. The latter is an ensemble classifier based on decision trees.

Problem 4.4

The neural network is described by the following parameters

- Two inputs
- One Hidden Layer with two units.
- Output layer with one unit.
  - Two neurons fully connected with input. The following matrices have the parameters of the units: number of rows is number of units and columns is number of inputs of the untis.
    - Weights
    $$\mathbf{W}_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$
    - and the bias vector
    $$\mathbf{b}_1 = \begin{pmatrix} 0.5 \\ -1.5 \end{pmatrix}$$
  - One output Layer with one neuron. Fully connected with hidden layer.
    - Weights $\mathbf{W}_2 = (0.7 \quad -0.4)$
    - bias $b_2 = -1$

- Activation functions: step function (sign function)for all neurons

$$o = sign(net) = \begin{cases} 1 & net \geq 0 \\ -1 & net < 0 \end{cases}$$

1. Represent the neural network and its parameters graphically.
2. Assume that the network is feed-forward and the inputs are the two attributes of the data set described in table 3. Compute the values of the output for the different examples. Sol.: ID1: y = 1, ID2: y = 1; ID3: y = -1, ID4: y = -1;
3. The neural network solve the classification problem?

ID	attribute 1 (x_1)	attribute 2 (x_2)	class	$\lambda_i d_i$
1	-1	1	A	-0.5
2	1	-1	A	0
3	-1	-1	B	0.5
4	1	1	B	0

Table 3: Toy data set: two features and two classes. The last column entries have the results of training the SVM (see problem 4.5).

Problem 4.5
The data set of table 3 was used to train a SVM with the following kernel function

$$k(\mathbf{x}, \mathbf{z}) = (\gamma \mathbf{x}^T \mathbf{z} + c)^d$$

assigning $\gamma = 1, c = 0, d = 2$ the last column of the table and $b = 0$ are the results of the training phase.

1. Is the data set linearly separable?
2. What is the numerical value assigned to class A? Sol.: -1
3. Compute the values of the decision rule for all the values of the training set.sol:$-1, -1, 1, 1$
4. Write a program to calculate the class of the points in a grid between -2 and 2 for both features.

Problem 4.6
The data set of table 3 can be used to construct a decision tree. Consider that the entropy is used to construct the tree.

1. Compute the entropy I of the root node.

2. Consider the decision rule for root node $x_1 > 0$ and calculate the decrease in impurity. Sol: $\Delta(I) = 0$

3. Complete the tree. Indicate the values of the decrease in impurity with the two decisions rules you need to complete the tree.

4. Indicate in the $2D$ feature space the regions of each class.

Problem 4.7

The two moon data set was used to train 3 classifiers. The three classifiers were initialized in the scikitlearn with the following parameters

mlp= MLPClassifier(activation='tanh', hidden_layer_sizes=(4,2), max_iter=5000)
svm=SVC(C=1.0,kernel='linear', max_iter=1000)
forest = RandomForestClassifier(max_depth=3, min_samples_split=5,n_estimators=100, max_features='log2', oob_score=False)

1. Draw the neural network corresponding to mlp.

2. Explain parameters C e *kernel* of svm.

3. How many trees has the classifier defined by forest. What is the meaning of the other parameters?

4. All the models can be initialized without the user assigning the hyperparameters. There are default values. Find out these values for each model.

5. The following figure shows the decision surfaces created by the training sessions of the three classifiers. Identify the classifier that creates each of the decision surfaces.

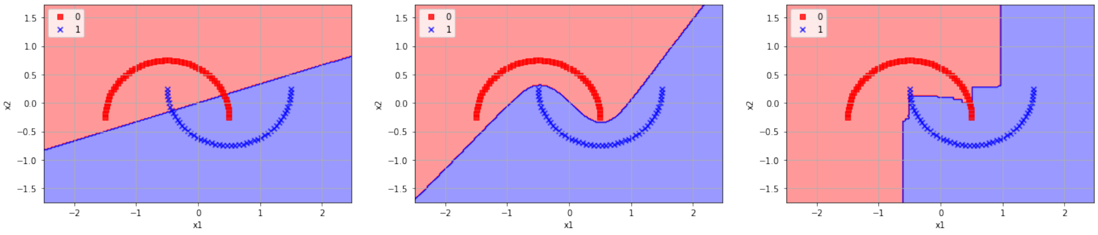


Figure 4.1: Decision surfaces after training the classifiers.

Problem 4.8
Compute the prediction (Play or Not Play Tennis) of the example

Day	Outlook	Temperature	Humidity	Wind
	rain	hot	High	Weak

using a Naive Bayes model trained with the data of Figure 1.1.

5 Evaluation of Classifiers

A set of measures are used to evaluate the classifier: accuracy, error rate, precision , recall and so on [3]. Usually as the amount of data is limited there also strategies to repeat the measures and consequently present the results with statistical confidence. After training (adaption of the parameters of the model) the classifier has to be evaluated with a test set. The results of classification in the test set are used to construct a confusion matrix.

Problem 5.1
The problem under analysis is related with USA elections and the contents of the following table is the confusion matrix [1].

	True(Actual) Class	
Predicted Class	Democrat	Republican
Democrat	81	2
Republican	6	46

1. Calculate the size of the data set. Sol.: SizeDS: 135

2. Calculate the following parameter: accuracy, error rate (misclassification error), recall and precision. Sol.: Acc = 94.07%, ErRate = 5.93%, Rec = 93.10%, Prec = 97.59%

Problem 5.2
The contents following table is the confusion matrix calculated to evaluate the performance of a model for the glass dataset [2] using cross-validation strategy.

	True Class					
Predicted Class Class	A	B	C	D	E	F
A	52	15	5	0	0	1
B	10	50	6	2	1	3
C	7	6	6	0	0	0
D	0	2	0	10	0	1
E	0	1	0	0	7	0
F	1	2	0	1	2	24

1. Calculate the size of the data set, the number of classes, the number of instances per class. Sol.: SizeDS: 215, #classes: 6, #inst_Classe: A:70, B:76, C:17, D:13, E:10, F:29

[2] http://archive.ics.uci.edu/ml/datasets/Glass+Identification

2. Calculate accuracy, precision and recall for each class. Sol.: Acc = 69.30%, Rec = 74.29%, Prec = 71.23%
3. Assuming that the classifier was evaluated with 5−fold cross-validation describe a possible solution for the contents and size of each partition if the sampling follows a stratified strategy.
4. Reading the description of the data set in the repository find out what is the problem addressed with this data set. How many features have the examples of the data set?

Problem 5.3

Two classifiers $Model_A$ and $Model_B$ where evaluated in $K = 10$ folds of a data set. The following table summarizes the results

	Accuracy (%)		
Fold	$Model_A$	$Model_B$	difference(d)
1	87.45	88.4	−0.95
2	86.5	88.1	−1.6
3	86.4	87.2	−0.8
4	86.8	86	0.8
5	87.8	87.6	0.2
6	86.6	86.4	0.2
7	87.3	87	0.3
8	87.2	87.4	−0.2
9	88	89	−1.0
10	85.8	87.2	−1.4
$m \pm \sigma$	86.99 ± 0.68	87.43 ± 0.89	-0.45 ± 0.81

How the two performances can be compared? Test the null hypothesis ($H_0 \rightarrow d \approx 0$).

References

[1] Max Bramer. Principles of Data Mining. Springer, 2007.

[2] Tom M. Mitchell. Machine Learning. McGraw-Hill, 1997.

[3] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. Introduction to Data Mining (2nd Edition). Pearson, 2nd edition, 2018.