Reconhecimento de Padrões/Inteligência Geoespacial/Aprendizagem Computacional em Biologia

2020/2021

Exame Normal 24 June 2021 Duration: 2h00

Name: **Number:**   **Practical Class:**

AVISO/WARNING

The Exam has a duration of 2h00m. The test is composed by five questions. The last question is a practical question in a language of your choice. Each question must be answered in the framed box below (and following) it. Questions may be answered in Portuguese or English. This is a closed book test. You may use only 1 A4 manuscript with your 'own' notes. You are allowed to use a calculator machine. Violation of the rules ends up with exam cancellation, course failure and eventually you may be subject to disciplinary procedure. If you have any questions, you may ask. Good Luck!

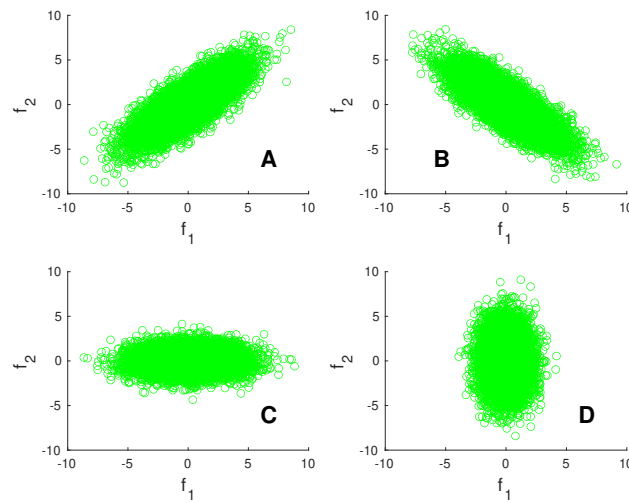| Question	pts	Results	Graded by:
1)	15		
2)	20		
3)	15		
4)	20		
5)	30		

Graded by:

Question 1 - PCA

☐ **15 pts**

Figure 1: Data Distributions

a) Relate, justifying your choices, the following covariance matrices with each one of the data distributions in Figure 1:

$$C_1 = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}$$

$$C_2 = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$$

$$C_3 = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$$

$$C_4 = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$$

b) What is the variance of f_1 and f_2 in each case? Justify why.

c) If PCA is applied how would you calculate the eigenvectors for each dataset? Justify.

Your answer to 1):

Question 2 - Bayes Decision Theory

☐ **20 pts**

Suppose we want to solve a classification problem with Bayes Decision Theory. Observe the following figure which illustrates the conditional probabilities for a two-class problem. The positive class is w_1 and the negative class is w_2.
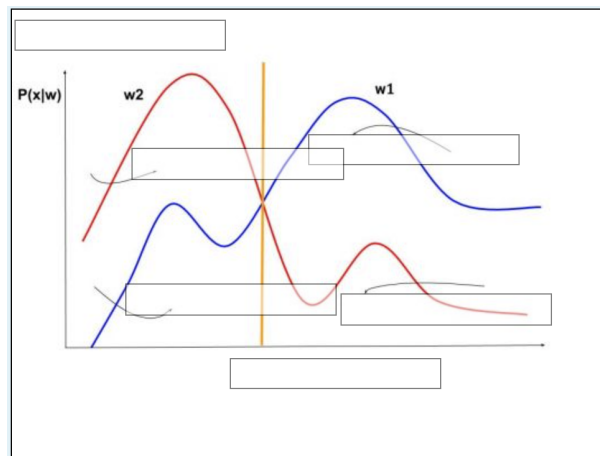

Figure 2: Joint probabilities for classes w_1 and w_2

(a) Fill in the white boxes appropriately (6 boxes).
(b) Give the Bayes Classification rule.
(c) In which point is the probability of error minimized. Justify.
(d) What happens if you shift the threshold to the right? Give an example where this situation would be seemingly favourable?

Your answer to 2):

Cont. your answer to 2):

Question 3 - Support Vector Machines

☐ **15 pts**

(a) Once we have found support vectors and the weights of a Support Vector Machine, how do we find a classification vector ? Justify.

(b) In a soft-margin support vector machine, decreasing the slack penalty term C causes: Select one or more:

1. more overfitting

2. a smaller margin

3. less overfitting

4. less sensitivity to outliers

(c) Draw a figure illustrative of the question above. Justify.

Your answer 3):

Cont. your answer to 3):

Question 4 - Hybrid Classifier

□ **20 pts** Figure 3 illustrates a basket with apples and tangerines with similar but distinct patterns. Suppose you have a dataset with similar images of each fruit.

Figure 3: Fruits Basket

You want to find a robust classifier that is able to distinguish between apples and tangerines. For that, you have tried 3 classifiers whose Posteriori Probabilities are indicated as illustrated in table 1.

Posteriori Probabilities

Classifier	Apples	Tangerines
Fisher LDA	0.93	0.88
KNN	0.60	0.70
Bayes	0.90	0.84

(a) From the experience in your PR course project what is the single algorithm listed in the table above of your preference and why. Give a brief description of the algorithm.
(b) Using a **hybrid classifier** calculate which class is found by this classifier by taking into account the Posteriori Probabilities obtained for each classifier. Illustrate all the steps you used in your calculations.
(c) Indicate which rule you used for the hybrid classifier. Justify.

Your answer 4):

Cont. your answer to 4):

Question 5 - kNN optimization

☐ **30 pts**
Write a function to search for the best K of a kNN classifier. The function should perform a grid-search looking for the appropriate parameter value for a given dataset. You can choose your language of choice. In Matlab the function should have the following prototype:

- **function [MeanF1,StdF1]=optimize_knn(Xtr,Ttr,Ks,n_runs)**

Where:

- **Xtr** is a matrix with dimensions D x Ptr, being D the problem dimensionality and Ptr the number of patterns in the training data;

- **Ttr** is the target vector with dimensions 1 x Ptr, and with "1" labeling positive patterns and "2" labeling negative patterns;

- **Ks** is a vector containing the K values to be considered.

- **n_runs** defines the number of runs. This parameter defines the number of times that grid-search should be performed aiming to have appropriate statistics about the influence of the different parameter combinations;

- **MeanF1** is a vector containing the average F1 measure values for each element in Ks.

- **StdF1** is a vector containing the standard deviation of the F1 measure values for each element in Ks.

Your answer to 5):

Cont. your answer to 5):