Question 1)

%Exercise 4 - Question 1

clear all

close all

load('t082');

a = size(x,2);

b = length(x);

%covariance of x

cv = cov(x);

%eigenvectors of cv, v & d

[v,d] = eig(cv);

%contributions of each eigenvector

f1 = d(1,1);

f2 = d(2,2);

f3 = d(3,3);

for farr = [f1 f2 f3]

pf = farr/(f1+f2+f3)

end

a) b) This script gives us the values of a and b, those numbers are the dimensionality of the feature space (a = 3) and the number of data samples (b = 100).

c) d) Then, we calculate the covariance of x and the eigenvectors of that covariance, so we can calculate the contributions of each eigenvector. By doing that, we obtain 2 new matrices that are v and d, the one that interest us the most for now is d because it has the values of the contributions of each one of the eigenvectors. Each eigenvector is related to one feature, so we can suppose that the contribution of the feature and the contribution of the eigenvector is going to be the same. The results are:

cv =

1.3512 -1.2631 -1.2837

-1.2631 2.1851 -0.0013

-1.2837 -0.0013 3.6328

v =

0.8152 0.3345 -0.4729

0.4970 -0.8232 0.2744

0.2975 0.4587 0.8373

d =

0.1127 0 0

0 2.6991 0

0 0 4.3573

Contributions of each eigenvector/feature:

1. pf = 0.0157
2. pf = 0.3765
3. pf = 0.6078

Question 2)

The vect\_arrow() function happened to be on the internet and its purpose is to draw the vectors. The scripts are attached files on the zip of the tasks. In

Question 3)

Assuming that my example case is a system designed to classify and differentiate apples from lemons:

* **“Into which subsets you would divide the available data for the selected case, and what is the purpose of each subset?”** – I would divide the available data on different subsets depending on what the data of a single fruit is about. The obvious thing to do is just have 2 different subsets for each feature measured. In that case we would have either a lemon or an apple as an output of measuring, depending on what parameters we assign to our measuring system in order to tell if the fruit is an apple or a lemon. The purpose of each subset is to have all of the fruit separated into 2 different classes. Even if there are some fruit damaged of some other failures, we can make sure that they are on one of those subsets and then, later on the process of selecting, we could have some other classification process or system that separates the good fruit to the one damaged or the one that does not interest us.
* **“How would you select the data samples into each subset?”** – When it comes to selecting data into subsets, we have to establish the threshold depending on what the features that we have selected. For example, the length of our fruits would be different if the fruit is a lemon or if it´s an apple. The value of the threshold is going to depend on what output data we receive from the system because is better to adapt to what outcomes we have received than creating a value before analyzing the data. No threshold is perfect and there are going to appear misclassifications for sure, but even with that in mind, we can make the system work as good as possible by analyzing the data and selecting the best possible features (those are going to be the ones that make a big difference between apples and lemons, for example the color of the fruit).
* **“Which are the relevant characteristics of the design task and the available dataset that affect your choices?”** – Related to my design task, the most relevant characteristics are: Feature selection (selecting the best features and looking for how we should apply our threshold for diving the data), Classification (divide the data) and Analysis (analyze the system and how it works and what works better for our case). In order to have a good system, we have to have in mind all of these and what should they be. According to the lectures the order should be: Data collection, Feature selection, Classifier design and System evaluation. I assumed that the data would be already in our system, that’s why I did not think about collecting it. For me, the most critical part is the analysis because it’s the one that is going to decide in the future if the system is going to work properly or not. Related to the available dataset, the most relevant characteristics are going to be: Extraction of the data and Possible errors and misclassifications. Extracting the data is easy but classifying it is not, so we should have some rules and thresholds for testing our model and approaching it to the best possible one, the one that has more efficiency when classifying both fruits. There is always going to be some errors that enter our system and we have to be able to notice them and be able to eliminate them from the system. The best decision-maker when it comes to errors is one that notices quick and fast changes in our data, for example a fruit that shouldn’t be there and has another color or even another length.

Question 4)

To measure the performance of a pattern recognition method we must determine how many classification errors occur for each class. For doing that, we are going to use the binomial distribution, and calculate the probability of error for each class. This measure is usually used to see the differences between different methods and see how they perform in a determined scenario. Usually is better to have more data to have a better classification error probability. The methods can be compared by measuring the error rate obtained when applying each one of them, then, the lowest error rate or the probability of error is going to determine the best method for our system.