UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

Advanced Engineering Data Analysis
MUEI, MUEA, and MASE
Prof: Daniel Fernández

Exercise 1

Topics: 2 problems about Principal Components Analysis and Linear Discriminant Analysis
Weight in Exercises grading: 50%

Notes:

- This activity must be done in groups (you are allocated into a group already)
- The report must have 10 sheets at most (5 pages). You can add appendices.
- Written in 11-12 font size.
- Submitted as a report (PDF or HTML format) to Atenea (under Exercise 1 folder).
- **Each member** of the group must submit the report (although it is the same).

The due date is at 23:59 on March 15, 2022

1. Principal Components Analysis (PCA)

The data file you are going to use for this exercise is **cars2004.csv**. This file is a comma-separated values file (CSV file). Thus, you can read it with the R command *read.csv* (i.e., *read.csv("cars2004.csv", header = TRUE)*).

The data consists of 425 cars from the 2004 model year and 19 features. The first feature is the name of the car (variable *Name*). Apart from that, seven features are binary indicators; the other 11 features are numerical (see Table 1).

Variable	Meaning
Sports	Binary indicator for being a sports car
SUV	Indicator for sports utility vehicle
Wagon	Indicator
Minivan	Indicator
Pickup	Indicator
AWD	Indicator for all-wheel drive
RWD	Indicator for rear-wheel drive
Retail	Suggested retail price (US$)
Dealer	Price to dealer (US$)
Engine	Engine size (liters)
Cylinders	Number of engine cylinders
Horsepower	Engine horsepower
CityMPG	City gas mileage
HighwayMPG	Highway gas mileage
Weight	Weight (pounds)
Wheelbase	Wheelbase (inches)
Length	Length (inches)
Width	Width (inches)

Table 1: Features for the 2004 cars data.

Apply a pre-processing of the data (Exploratory data analysis, missings, outliers, etc.) and apply PCA as we saw in class. You must justify all the steps you take.
You must also contextualize all results (i.e., what are the more important results that you observed? how do you interpret each principal component? do you have suggestions? how do you interpret the biplot? etc.)

2. Linear Discriminant Analysis and extensions

The data file you are going to use for this exercise is **phoneme.csv**. This file is a comma-separated values file (CSV file). Thus, you can read it with the R command *read.csv* (i.e., *read.csv("phoneme.csv", header = TRUE)*).

The data set contains samples of digitized speech for five phonemes: aa (as the vowel in *dark*), *ao* (as the first vowel in *water*), *dcl* (as in *dark*), *iy* (as the vowel in *she*), and *sh* (as in *she*). In total, 4509 speech frames of 32 msec were selected. For each speech frame, a log-periodogram of length 256 was computed, on whose basis we want to perform speech recognition. The 256 columns labeled *x.1* to *x.256* identify the speech features, while the columns *g* and *speaker* indicate the phonemes (labels) and speakers, respectively. Use only the first 10 columns, i.e., from x.1 to x.10, and the labels (column *g*).

Apply a complete pre-processing of the data (exploratory data analysis, missings, outliers, scaling, etc.). Apply a Linear, Quadratic, Regularized, and Mixture Discriminant Analysis as we saw in class (method, prediction, plotting,…). Which gives the best result? You must justify all the steps you take and contextualize all results.
Finally, print the posterior probabilities from the test data set for the first 6 rows (Note: that was not shown in class, so you should figure out how to compute them).