Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
January 17th, 2019

First Name and Family Name:
ID Number:

Problem n.1

File `luggage.txt` collects the weight [kg] of the pieces of luggage checked-in by 232 couples for the outbound and return flights of the route Milan-New York. Assume the observations to be a sample from a four-variate Gaussian distribution.

a) Perform a statistical test to verify if a significant difference exists between the mean weight of the pieces of luggage checked-in by men and by women.

b) Perform a statistical test to verify if a significant difference exists between the mean weight of the pieces of luggage checked-in in the outbound flight and in the return flight.

c) Provide four T^2 simultaneous confidence intervals (global confidence 90%) for appropriate mean differences of weight to support the conclusions obtained at point (a) and (b). Comment the results.

d) Knowing that the maximum allowance for checked-in luggage is 23 kg, would you believe (with probability 90%) that the next female passenger on the flight New York - Milan will be charged extra-weight for her luggage?

Problem n.2

A geologist has recently analyzed a collection of 335 diamonds randomly taken in the *Fortune diamond mine* in South Africa. The file `diamonds.txt` contains the diameter (mm) and weight (carats) of those diamonds. The geologist supposes the existence of different types of diamonds. After having standardized the variables and by using an agglomerative hierarchical clustering algorithm (based on Euclidean distance and Ward linkage):

a) Provide a guess on the possible number of diamonds types and on their percentages in the sample.

b) Using Bonferroni's intervals (with global confidence level of 95%) detect the differences among the different types of gems.

Problema 3

The dataset `trading.txt` reports the 252 observations of the increments between consecutive days of 2015 of the value of stocks of Google and Apple (i.e., the i-th observation represents the difference between the value at day d_i and the value at day $d_i - 1$). The dataset reports a further variable 'gain', that represents whether the value of Facebook's stocks had or not a positive increment the day after (i.e., the i-th observation of 'gain' is 'gain'=1 if the increment of the Facebook's stock value was positive between the day d_i and $d_i + 1$, 0 otherwise).

a) Build a classifier for the variable 'gain' based on the increments of Google and Apple stock values, by using quadratic discriminant analysis. Report the mean within the groups, the prior probabilities estimated from the sample and a qualitative plot of the regions of classification.

b) Estimate the AER of the classifier through the APER and by leave-one-out cross-validation. Comments the results, with particular reference to the comparison with the trivial classifier.

c) Would you sell your Facebook stocks if between yesterday and today there was an increment of $(-3, -1.2)'$ in the values of Google's and Apple's stocks?

Problem n.4

The file `areaC.txt` collects the data on the number of accesses in the Area C of the Municipality of Milan, observed in March 2016. In particular, the reported data refer to the number of accesses for some types of vehicles (petrol, diesel, electric, GPL, natural gas and hybrid vehicles), together with the total number accesses. In addition, the data include the indication about the type of day (variable: 'weekend').

a) Formulate a linear regression model for the total accesses, as a function of all the other variables. Include in the model a possible dependence of the total accesses on the variable 'weekend', but *only in the intercept*. Report the model and its parametrization, together with the estimates of the 10 parameters of the model (the coefficients $\beta_0, ..., \beta_9$ and the errors' standard deviation σ).

b) Analyse the residuals of the model. Highlight possible weaknesses of the model.

c) Perform a principal component regression to estimate the parameters of the model (a). Interpret the principal components. Report the new parametrization of the model and the estimates of the model parameters. Evaluate the new diagnostic plots.

d) Propose a reduction of model (c) by performing appropriate statistical test(s) and update the model parameters.