Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
July 17th, 2020

Problem n.1

In order to design a new smart-office sensing device, a study is conducted aiming at detecting the occupancy of an office room from environmental measurements. The file `occupancy.txt` reports the measurements of relative humidity (in percentage) and CO2 concentration (in hundreds of ppm) collected in an office room at 100 time instants randomly distributed over the 24 hours of a working day. Moreover, a binary label indicating the occupancy of the room (0 for not occupied, 1 for occupied status) is reported in the file.

a) Assuming that the office room is occupied on average 9 hours a day, build a classifier for the room occupancy that minimizes the expected number of misclassifications. Specify the model and verify the model assumptions. Report the estimates of the model parameters and a plot of the classification regions.

b) Compute the APER of the classifier.

c) Use the classifier to classify a new measurement with 26% of humidity and 900ppm of CO2.

d) Build a k-nearest neighbor classifier for the room occupancy choosing the parameter k equal to 5. Report a plot of the classification regions and compute the error rate on the training set. Compare the results with points (a) and (b) and comment on the performances of the classifiers.

Upload your results here:
https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUNzVHQ1hF

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
July 17th, 2020

Problem n.2

A jeweler is considering changing diamond supplier in order to purchase diamonds of superior quality. The jeweler performed identical and independent measurements on 8 diamonds provided by the current supplier and on 8 diamonds provided by the new supplier in order to compare the quality of the diamonds. The file `purity.txt` contains the values of 10 purity parameters measured by the jeweler on the 16 diamonds. The higher the purity parameter, the higher the quality of the diamond.

a) For each purity parameter, perform a permutation one-sided test to look for possible statistical superiority of the diamonds of the new supplier. In detail, for each purity parameter, use the difference of the sample means as test statistic and use 5000 random permutations with random seed equal to 123 to estimate the permutational distribution. Report the value of the 10 test statistics and their corresponding p-values.

b) For which purity parameters the new supplier can be considered superior to the current supplier if the jeweler wants to limit the false discovery rate to a maximum value of 10%?

c) For which purity parameters the new supplier can be considered superior to the current supplier if the jeweler wants to impose a probability at most 1% that at least one of the non-superior purity parameter is judged as superior?

Upload your results here:
https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglURFlaVjY1OU

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
July 17th, 2020

Problem n.3

Different chemicals are tested to predict aquatic toxicity towards Daphnia Magna, a small planktonic crustacean. The file `toxicity.txt` contains the values of 6 molecular descriptors (C1,...,C6) of the 100 tested chemicals and a measure of toxicity (tox), which is the concentration that causes death in 50% of test population of Daphnia Magna over a test duration of 48 hours.

a) Formulate a linear regression model for the toxicity, as a function of all the other variables. Report the estimates of the parameters and verify the assumptions of the model.

b) Predict the toxicity of a new chemical characterized by the following molecular descriptors: C1=100, C2=0.7, C3=2, C4=4, C5=1.4, C6=3. Provide a pointwise estimate and an interval of level 95%.

c) Perform a variable selection through a Lasso method, by optimizing via cross-validation the parameter controlling the penalization (λ) within the range [0.01; 1]. Report the optimal λ and the significant coefficients.

d) Answer point (b) using the reduced model obtained at point (c).

Upload your results here:
https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUMjRNMExK

Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
July 17th, 2020

Problem n.4

The file `traffic.txt` contains hourly measurements of traffic volume on highway I-94 near Minneapolis recorded for 30 days. Consider a functional data analysis approach where, for each day, the measurements provided are considered as discrete sampling of underlying smooth functions.

a) Perform a smoothing of each daily data through a projection over a B-spline basis with 15 basis elements of degree 3. Provide a plot of the smoothed data and report the first 3 coefficients obtained for Day 1.

b) Perform a functional principal component analysis of the smoothed data obtained at point (a). Report the variance explained along the first 3 functional principal components, the screeplot and a plot of the first 3 eigenfunctions.

c) Propose a possible dimensionality reduction for the data, interpret the retained principal components and discuss the results.

d) Provide a plot of the scores along the first functional principal component as a function of the day number. Discuss the result to further enhance the interpretation.

Upload your results here:
https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUOTNURTNK