Statistical Methods for Data Science

DATA7202

Semester 1, 2021

Assignment 2 (Weight: 25%)

Assignment 2 is due on 19 Apr 2021 17:00).

Please answer the questions below. For theoretical questions, you should present rigorous proofs and appropriate explanations. Your report should be visually appealing and all questions should be answered in the order of their appearance. For programming questions, you should present your analysis of data using Python, Matlab, or R, as a short report, clearly answering the objectives and justifying the modeling (and hence statistical analysis) choices you make, as well as discussing your conclusions. Do not include excessive amounts of output in your reports. All the code should be copied into the appendix and the sources should be packaged separately and submitted on the blackboard in a zipped folder with the name:

```
"student_last_name.student_first_name.student_id.zip".
```

For example, suppose that the student name is John Smith and the student ID is 123456789. Then, the zipped file name will be `John.Smith.123456789.zip`.

1. **[15 Marks (see details below)]** Consider the Hitters data-set from Assignment 1 (given in Hitters.csv) and recall that our objective is to predict a hitter's salary via linear models.

   (a) **[10 Marks])** Apply Principal Component Regression (PCR) with all possible number of principal components. Using the 10-Fold Cross-Validation, plot the mean squared error as a function of the number of components and determine the optimal number of components.

   (b) **[5 Marks]** Apply the Lasso method and plot the 10-Fold Cross-Validation mean squared error as a function of λ. Determine the best λ and the corresponding mean squared error.

2. **[15 Marks]** Consider the data given in `ships.csv`. There are 34 observations that contain a ship type (coded 1-5 for A, B, C, D and E), year of construction (1=1960-64, 2=1965-70, 3=1970-74, 4=1975-79), period of operation (1=1960-74, 2=1975-79), months of service (63 to 20,370), and the response variable damage incidents, which ranges from 0 to 53.

   Construct a Poisson regression model and report the coefficients (for type, construction, operation, and months), and the corresponding 95% CIs. You can use the `statsmodels.api` module.

3. **[30 Marks (see details below)]** A soft drink bottler is analyzing vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. This service activity includes stocking the machine with beverage products and minor maintenance or housekeeping. The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time are the number of cases of product stocked and the distance walked by the route driver. The engineer has collected 25 observations on delivery time (minutes), number of cases and distance walked (feet). The data is in the file "softdrink.csv".

   (a) **[10 Marks** Compute the multiple regression of Time on Cases and Distance. State the fitted model, the estimated residual standard deviation, and the P-values for the overall model and each of the two predictors.

(b) **[10 Marks** Obtain residual plots and the histogram of the residuals. Comment on these.

(c) **[10 Marks** There is an observation in this data set which is extremely influential according to Cook's distance. Which observation is it? Display a Cook's distance plot to determine the Cook's distance of the next most influential observation.

4. **[15 Marks]** Conjugate Uniform random variable analysis: Consider n iid Uniform random variables Y_i, $(i = 1, \ldots, n)$, each with p.d.f.

$$\mathbb{P}(y \mid \theta) = \frac{1}{\theta}.$$

Suppose that the prior for θ is the Pareto distribution $\mathsf{Pareto}(\alpha, x_m)$. Namely

$$p(x \mid \alpha, x_m) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}} & x \geqslant x_m \\ 0 & x < x_m. \end{cases}.$$

Derive the posterior distribution of θ.

5. **[25 Marks (see details below)]** Consider a sampling from the 2-dimensional pdf

$$f(x, y) = c\, \mathrm{e}^{-(xy+x+y)}, \quad x \geqslant 0, \quad y \geqslant 0,$$

for some normalization constant c, using a Gibbs sampler. Let $(X, Y) \sim f$.

(a) **[10 Marks]** Find the conditional pdf of X given $Y = y$, and the conditional pdf of Y given $X = x$.

(b) **[15 Marks]** Write working Python code that implements the Gibbs sampler and outputs 1000 points that are approximately distributed according to f.