Statistical Methods for Data Science

DATA7202

Semester 1, 2024

Assignment 1 (Weight: 25%)

Assignment 1 is due on Wednesday 20.03.24 16:00).

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]** Repeat the advertisement exercise with the following changes.

   (a) The data is generated via the following data generation mechanism: $X_i \sim \mathsf{Gamma}(1,1)$ for $i \in \{1,2,3\}$; here $\mathsf{Gamma}(1,1)$ stands for the continuous Gamma distribution with both scale and shape parameters equal to 1.

   (b) In addition, the model for y is as follow:

   $$Y = 0.5X_1 + 3X_2 + 5X_3 + 5X_2X_3 + 2X_1X_2X_3 + W, \tag{1}$$

   where $W \sim \mathsf{N}(0,\sigma^2)$ where $\sigma = 2$.

   Similar to the original example, generate train and test sets of size $N = 1000$. Fit the linear regression and the random forest models to the data. For the linear regression, make an inference about the coefficients, specifically, comment about the contributions of different advertisement types to sales. Use the linear model and the RF (with 500 trees), to make a prediction (using the test set), and report the corresponding mean squared errors.

   When constructing datasets, please use "1" and "2" seeds for the train and the test sets, respectively.

2. **[10 Marks]** Consider a function

   $$f(x) = 3 + x^2 - 2\sin(x) \quad 1 \leqslant x \leqslant 8.$$

   Write a Crude Monte Carlo algorithm for the estimation of

   $$\ell = \int_1^8 f(x)\,\mathrm{d}x,$$

   using $N = 10000$ sample size. Deliver the 95% confidence interval. Compare the obtained estimation with the true value ℓ.

3. **[10 Marks]** Consider the following variant of the cross-validation procedure.

   (i) Using the available data, find a subset of "good" predictors that show correlation with the response variable.

   (ii) Using these predictors, construct a model (for regression or classification).

   (iii) Use cross-validation to estimate the model prediction error.

   Is this a good method? Do you expect to obtain the true prediction error? Explain your answer. Please note that no coding is required here and one paragraph general answer is sufficient.

4. **[5 Marks]** Suppose that we observe $X_1, \dots, X_n \sim F$. We model F as a Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$. For this problem, determine the hypothesis class

$$\mathcal{H} = \{f(\mathbf{x}, \theta);\ \theta \in \Theta\}.$$

   and state explicitly what is θ and Θ.

5. **[15 Marks]** Let $\mathcal{H}$ be a class of binary classifiers over a set $\mathcal{Z}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$, and let g be a target hypothesis in $\mathcal{H}$. Show that the expected value of $\text{Loss}_{\mathcal{T}}(g)$ over the choice of $\mathcal{T}$ equals $\text{Loss}_{\mathcal{D}}(g)$, namely,

$$\mathbb{E}_{\mathcal{T}} \text{Loss}_{\mathcal{T}}(g) = \text{Loss}_{\mathcal{D}}(g).$$

6. **[15 Marks (see details below)]** Consider the following dataset.

x_1	y
0	1
1	2
2	3
3	2
4	1

   Now, suppose that we would like to consider two models.

$$\text{Model}_1: \quad y = \beta_0 + \varepsilon,$$

   and

$$\text{Model}_2: \quad y = \beta_1 x_1 + \varepsilon,$$

   where $\varepsilon \sim \mathsf{N}(0, 1)$. That is, we consider two linear models Model_1 is the constant model and Model_2 is a regular linear model without the intercept.

   (a) **[5 Marks)]** Fit these models tot the data and write the corresponding coefficients. Namely, fill the following table:

Model	β_0	β_1
Model_1		0
Model_2		

   (b) **[5 Marks)]** Consider the squared error loss, the absolute error loss, and the $L_{1.5}$ loss. Find the average loss for each model. Namely, fill the following table:

Model	squared error loss	absolute error loss	$L_{1.5}$ loss
Model_1			
Model_2			

(c) **[5 Marks)]** Draw a conclusion from the obtained results.

7. **[30 Marks (see details below)]** Consider as data-set given in data.csv. Our objective is to predict a response variable y via a linear model. The data contains 4 explanatory variables $(x1, x2, x3, x4)$. Here, $x1, x3$ and $x4$ are numeric variables and $x2$ is a categorical variable.

   (a) **[10 Marks)]** Load the data-set and replace all categorical values ($x2$) with numbers. You do not know if $x2$ contains an ordinary categorical variable or not, unfortunately. Therefore, you decide to create two datasets, where in the first dataset, $x2$ is considered to be ordinary, and, in the second dataset, $x2$ is assumed to be unordered.

   (b) **[20 Marks)]** Fit linear regression and report 10-Fold Cross-Validation mean squared errors for to datasets. What is your conclusion about $x2$, is it ordered or not?