Statistical Methods for Data Science

DATA7202

Semester 1, 2024

Assignment 2 (Weight: 25%)

Assignment 2 is due on Wednesday 17.04.24 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. **[10 Marks]** Suppose that in a certain leaf node of a decision tree that was applied to a classification problem, there are 3 blue and 2 red data points in a certain tree region. Calculate the misclassification impurity, the Gini impurity, and the entropy impurity. Repeat these calculations for 2 blue and 3 red data points.

2. **[5 Marks]** Let τ be a training set with n elements. Recall the bootstrap procedure that was discussed in class and consider the out of bag (OOB) error estimator. Show that a given bootstraped dataset τ^* and the corresponding tree constructed using this set τ^*, can be used to calculate an OOB error for about $0.37n$ data points of the original dataset τ.

3. **[10 Marks]** Consider the following train/test split of the data.

```
import numpy as np
from sklearn.datasets import make_friedman1
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import BaggingRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score

# create regression problem
n_points = 1000 # points
x, y = make_friedman1 ( n_samples =n_points , n_features =10 ,
noise = 5 , random_state =100)
# split to train /test set
x_train , x_test , y_train , y_test = \
train_test_split (x, y, test_size =0.33 , random_state =100)
```

Construct a bagging regressor with 100 trees and identify the optimal parameter m in the sense of R^2 score. Here, m is the subset size of predictors that are being considered at each split.

4. **[10 Marks]** Consider the following classification data and module imports:

```
from sklearn.datasets import make_blobs
from sklearn.metrics import zero_one_loss
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingClassifier

if __name__ == "__main__":
    x, y =  make_blobs(n_samples=1000, n_features=20, centers=2,
                           random_state=100, cluster_std=6)

    x_train , x_test , y_train , y_test = \
        train_test_split (x, y, test_size =0.33 , random_state =10)
```

Using the gradient boosting algorithm with $B = 80$ rounds, train the gradient boosting classifier using the train data and different values of γ, for $\gamma = 0.1, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$. Use the test data to determine the best value of γ with respect to zero-one loss.

5. **[15 Marks (see details below)]** Consider the Hitters data-set (given in Hitters.csv). Our objective is to predict a hitter's salary via linear models. Our objective is to predict a hitter's salary via linear models.

   (a) **[2 Marks)]** Load the data-set and replace all categorical values with numbers. (You can use the LabelEncoder object in Python).

   (b) **[3 Marks)]** Generally, as shown in Assignment 1, it is better to use dummy variables when dealing with categorical data. Justify the usage of LabelEncoder in (a).

   (c) **[5 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.

   (d) **[5 Marks]** Apply the Lasso method and plot the 10-Fold Cross-Validation mean squared error as a function of $\lambda \in \{10, 20, \ldots, 1000\}$. Determine the best λ and the corresponding mean squared error.

6. **[20 Marks (see details below)]** 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.

   (a) **[5 Marks]** 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.

   (b) **[15 Marks]** Using 1000 bootstrapped sampled datasets, construct the standard error and normal 95% CIs for the model coefficients. Is there a difference between the CIs?

7. **[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.