Statistical Methods for Data Science

DATA7202

Semester 1, 2021

Assignment 3 (Weight: 25%)

Assignment 3 is due on 17 May 21 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. **[10 Marks]** Show that any training set $\tau = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ can be fitted via a tree with zero training loss.

2. **[10 Marks]** Suppose during the construction of a decision tree we wish to specify a constant regional prediction function h^w on the region R_w, based on the training data in R_w, say $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\}$. Show that $h^w(\mathbf{x}) := k^{-1} \sum_{i=1}^{k} y_i$ minimizes the squared-error loss.

3. **[5 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.

4. **[15 Marks]** Suppose τ is a training set with n elements and τ^*, also of size n, is obtained from τ by bootstrapping; that is, resampling with replacement. Show that for large n, τ^* does not contain a fraction of about $\mathrm{e}^{-1} \approx 0.37$ of the points from τ.

5. **[30 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.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 =15 ,
noise =1.0 , 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 random forest regressor with 1000 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.

6. **[30 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
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import GradientBoostingClassifier

X_train, y_train =  make_blobs(n_samples=1000, n_features=10, centers=3,
random_state=10, cluster_std=5)
```

Using the gradient boosting algorithm with $B = 150$ rounds, plot the training loss as a function of γ, for $\gamma = 0.1, 0.3, 0.5, 0.7, 1$. What is your conclusion regarding the relation between B and γ?