ST4248 STATISTICAL LEARNING II

Midterm Examination (Semester 2 AY 2023/2024)

Name: _____

Matriculation No: _____

INSTRUCTIONS TO STUDENTS

1. There are **FOUR (4)** questions. Please answer all of them.
2. Upload a PDF copy of your solution through Canvas > Assignments> Midterm Submission by **2PM** on **MONDAY, 18 MARCH 2024**.
3. You are **NOT** to collaborate with your classmates for any part of the solution.

1. Consider a dataset $\{(\boldsymbol{x}_i, y_i) : 1 \leq i \leq n\}$ with predictor variables $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ and responses y_i. Assume the dataset is generated from the model

$$y_i = x_{i1}x_{i2} + \epsilon_i, \tag{1}$$

where ϵ_i are independent with mean 0 and variance σ^2. Boosting is applied to the dataset with d splits.

   (a) (5 marks) Describe the boosting algorithm for regression trees.

   (b) (5 marks) Is the selection of $d = 1$ appropriate? How about the selection of $d = 2$ and $d = 3$? Explain.

   (c) (10 marks) Design and implement a Monte Carlo study to investigate your response to (b), with the following components:

      (i) Generate n observations from model (1).

      (ii) Create a training set consisting of the first $n/2$ observations, and a test set consisting of the remaining observations.

      (iii) Perform boosting on the training set with $d = 1$, 2 or 3 splits (while keeping all the other parameters fixed).

      (iv) Produce a plot with d on the x-axis and corresponding test set MSE on the y-axis.

      Please append a printout of your R code to the solution.

   (d) (5 marks) Do you expect the training MSE to decrease as the number of trees B increases? You may assume that all the other parameters remain the same as B increases.

2. The `Carseats` dataset is available in the library `ISLR`. Treat the feature `Sales` as a quantitative response and the other features as predictor variables. Answer the following questions and append a printout of your R code to the solution.

   (a) (5 marks) Split the dataset equally into training and test subsets. Fit a regression tree with default parameter values to the training set. Plot the tree and interpret the results. What test MSE do you obtain?

   (b) (5 marks) Prune the tree from (a) down to k terminal nodes (leaves) for all possible values of $k \geq 2$. Plot the test MSE against the number of terminal nodes. Provide some comments on the plot.

   (c) (5 marks) Use bagging to analyze this dataset. What is the test MSE? Use the `importance` function to determine which variables are most important.

   (d) (5 marks) Use random forests to analyze this dataset. Do not specify `mtry`. What `mtry` is used by `randomForest`? What test MSE do you obtain? Use the importance function to determine which variables are most important. What can you say about the importance of the variables in bagging and random forest?

   (e) (5 marks) Find the test MSE for boosting with $d = 2$ splits. You can use the default shrinking parameter and number of trees. Provide some comments.

3. Consider the dataset $\{(x_i, y_i) : 1 \leq i \leq 100\}$ in `Q3.csv`, which has been generated from the model

$$y_i = m(x_i) + \epsilon_i, \tag{2}$$

where ϵ_i are independent with mean 0 and variance σ^2. To estimate $m(x)$, we approximate it by the cubic spline with 3 knots,

$$\begin{aligned} m(x) \approx & \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\ & + \beta_4 (x - 0.3)_+^3 + \beta_5 (x - 0.5)_+^3 + \beta_6 (x - 0.7)_+^3. \end{aligned}$$

Answer the following questions and append a printout of your R code to the solution.

(a) (5 marks) Fit the model and write down the estimated function $\widehat{m}(x)$

(b) (5 marks) To investigate whether $m(x)$ indeed varies with x, test the following hypothesis at significance level 0.05,

$$H_0 : m(x) \equiv \text{a constant}.$$

(c) (5 marks) For the sequence of x values $0, 0.1, 0.2, ..., 0.8, 0.9, 1$, predict the expected y values and obtain the corresponding 95% confidence intervals.

(d) (5 marks) Suppose $m(x)$ is linear in the boundary regions $x < 0.3$ and $x > 0.7$. Show that this implies the constraints $\beta_2 = 0$, $\beta_3 = 0$, $\beta_4 + \beta_5 + \beta_6 = 0$ and $0.3\beta_4 + 0.5\beta_5 + 0.7\beta_6 = 0$.

(e) (5 marks) How many degrees of freedom does the constrained cubic spline in (d) have?

4. Answer the following questions and append a printout of your R code to the solution.

   (a) (5 marks) Generate a simulated two-class data set $\{(\boldsymbol{x}_i, y_i) : 1 \leq i \leq 100\}$ with features $\boldsymbol{x}_i = (x_{i1}, x_{i2})$ and binary responses y_i, in which there is a visible but non-linear separation between the two classes. Split the dataset equally into training and test subsets.

   (b) (5 marks) Fit a support vector classifier on the training data, using CV to choose `cost`. What is the test error rate?

   (c) (5 marks) The decision boundary in (b) is given by $f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. In class, an alternative representation is given as

   $$f(\boldsymbol{x}) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle,$$

   where $\mathcal{S}$ is the collection of indices of the support vectors. Express β_1 and β_2 in terms of α_i and $\boldsymbol{x}_i$, $1 \leq i \leq 100$.

   (d) (5 marks) Fit a support vector machine on the training data with a polynomial kernel (with degree $d > 1$), using CV to choose `cost`. Compare the test error rate with (b).

   (e) (5 marks) For your choice of degree d in (d), derive the transformed features $h(\boldsymbol{x}) = (h_1(\boldsymbol{x}), ..., h_p(\boldsymbol{x}))$ such that the decision boundary has the form $f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j h_j(\boldsymbol{x})$.

END OF PAPER