Elements of Machine Learning, WS 2023/2024
Jilles Vreeken and Krikamol Muandet
Homework Sheet #6: *Trees and Forests and SVMs*

Deadline: Thursday, February 01, 2024, 15:00

Before solving the exercises, read the instructions on the course website.

- For each theoretical problem, submit a single `pdf` file that contains your answer to the respective problem. This file may be a scan of your (legible) handwriting.
- For each practical problem, submit a single `zip` file that contains
  - the completed jupyter notebook (`.ipynb`) file,
  - any necessary files required to reproduce your results, and
  - a `pdf` report generated from the jupyter notebook that shows all your results.
- For the bonus question, submit a single `zip file` that contains
  - a `pdf` file that includes your answers to the theoretical part,
  - the completed jupyter notebook (`.ipynb`) file for the practical component,
  - any necessary files required to reproduce your results, and
  - a `pdf` report generated from the jupyter notebook that shows your results.
- Every team member has to submit a signed Code of Conduct.
- **IMPORTANT** You must make the team on CMS *before* you upload the solutions. If you upload the solutions first and create the team after it, the solution will not show for the new team member!

Problem 1 (T, 3 Points). **Interpretibility**
In the lecture, you have seen the term "interpretability" come up to describe certain models.

1. (1 Point) Describe what we mean by *interpretability*.
2. (1 Point) "An interpretable model is *always* a better option than a less interpretable one". Explain whether you agree or disagree with this statement.
3. (1 Point) We have learned in the lecture that Trees are among highly interpretable models. Suppose you learn that Neural Networks, a highly *non*-interpretable model class, can always be represented by a decision tree. Does that mean Neural Networks are now highly interpretable? Explain whether you agree or disagree.

Problem 2 (T, 8 Points). **Decision Trees**

1. (3 Points) Sketch a tree corresponding to the partition of the predictor space indicated in Fig. 1. The numbers inside the boxes indicate the mean of Y within each region.
2. (3 Points) Now look at the tree shown in Fig 2. Sketch what the partition space for this tree would look like.
3. (2 Points) Consider four cases (a-d) shown in Fig 3. Each case represents a decision boundary between two classes marked in green and blue respectively. For each of the cases, choose if the decision boundary is better modelled by linear-regression or a classification tree, or neither of the two. Justify your choice in each case.

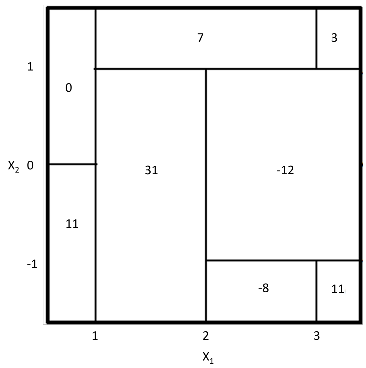

Figure 1: Predictor Space over variables X_1 and X_2

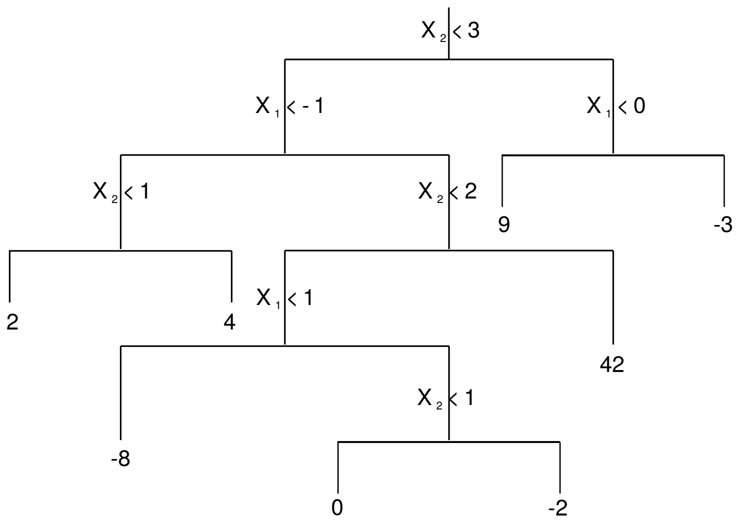

Figure 2: Decision tree over variables X_1 and X_2

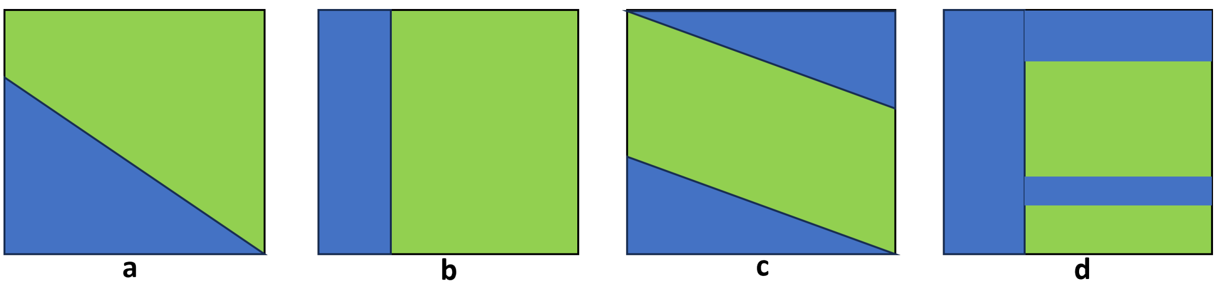

Figure 3: Decision boundaries defined between two classes (blue and green)

Problem 3 (T, 9 Points). **Support Vector Machines**

1. (3 Points, ISLP Exc. 9.7.2) We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$. We now investigate a non-linear decision boundary.
   (a) (0.5 Point) Sketch the curve $(1 + X_1)^2 + (2 - X_2)^2 = 4$.
   (b) (0.5 Point) On your sketch, indicate the set of points for which $(1 + X_1)^2 + (2 - X_2)^2 > 4$ and the set of points for which $(1 + X_1)^2 + (2 - X_2)^2 \leq 4$
   (c) (1 Point) Suppose that a classifier assigns an observation to the blue class if $(1 + X_1)^2 + (2 - X_2)^2 > 4$, and to the red class otherwise. To what class are the following observations classified? $(0, 0), (-1, 1), (2, 2), (3, 8)$.
   (d) (1 Point) Argue that while the decision boundary in (c) is not linear in terms of X_1 and X_2, it is linear in terms of X_1, X_2, X_1^2, and X_2^2.

2. (2 Points) Provide and explain one case where you would prefer *One versus One* approach to *One vs All* for multi-class classification. Provide one example and explain where you would prefer *One vs All*.

3. (4 Points) Consider the following optimization problem for Support Vector Classifier given in ISLP book Section 9.2.2. This classifier can be used to learn a linear decision boundary between two classes.

$$\text{maximize}_{\beta_0, \beta_1 \ldots, \beta_p, \epsilon_0, \epsilon_1, \ldots, \epsilon_p, M} M \tag{3.1}$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \; = \; 1, \tag{3.2}$$

$$y_i \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \right) \geq M(1 - \epsilon_i) \tag{3.3}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C. \tag{3.4}$$

   (a) (1 Point) Explain how the variables ϵ_i and C are related.
   (b) (1 Point) Explain why the following sentence is true or false: *"As the value of C increases, the bias of the classifier decreases and the variance increases"*
   (c) (1 Point) Write down the optimization problem if we want to change the linear decision boundary to a quadratic decision boundary.
   (d) (1 Point) How would you change the support vector classifier such that it can fit arbitrary non-linear decision boundary? *(Hint: Read Section 9.3 ISLP)*

Problem 4 (P, 15 Points). **Trees and Forests and Bags and Correlations**
In this exercise, we will study how correlated the individual trees in Bagging and in Random Forests are.

(a) (1 Point) Load the data in `train.csv` and compute the correlation of each predictor variable X_i with the target variable y.

(b) (2 Points) Use bagging to train $B = 50$ Regression Trees T^b on the data X.

(c) (3 Points) Load the data `test.csv` and compute the average correlation between the predictions $y^b_{\text{pred}} = T^b(X_{\text{test}})$ for different Trees $T^b, b = 1, \ldots, 50$.

(d) (2 Points) Similarly compute the average correlation between the residuals $y_{\text{test}} - y^b_{\text{pred}}$. Contrast the result with that derived from part (c) and explain which measure of correlation is more useful.

(e) (2 Points) For each $q \in \{0.25, 0.50, 0.75, 1.0\}$, train a Random Forest Regression with 50 trees on the training data, in which each tree uses only a fraction q of all available predictors.

(f) (2 Points) Recompute the correlations from (d) for each of the random forests, and plot the results in a suitable manner. Explain what you see.

(g) (3 Points) Compute the variable importances for each variable for each forest trained in (e). Plot them against the correlations computed in (a). What do you see? Why?

Note: All relevant models can be fit using `sklearn.ensemble.RandomForestRegressor`.

Problem 5 (Bonus, 10 Points). **Non axis-aligned Decision Trees**
In this problem, we will try to derive a method to learn non-axis aligned classification trees.

1. (1 Point) Look at the tree shown in Fig 4. Sketch the partition space for this classification tree.

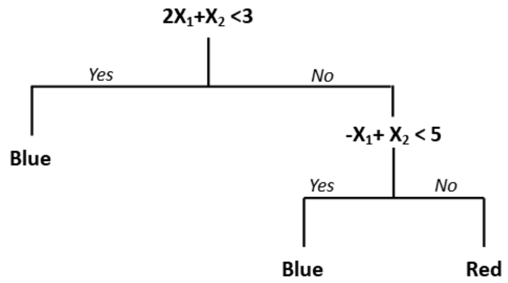

Figure 4: Predictor Space over variables X_1 and X_2

2. (1 Point) Now Sketch a tree corresponding to the partition of the predictor space indicated in Fig. 5. Names inside the regions indicate the predicted class for that region.

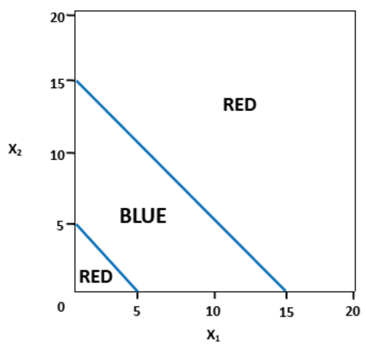

Figure 5: Decision tree over variables X_1 and X_2

3. (0.5 x 2 Points) Recall from the lecture that for axis-aligned splits, we had the split rule of the form $R_1(j, s) = \{X|X_j < s\}$ and $R_2(j, s) = \{X|X_j \geq s\}$. What would be the split rule for the cases shown in parts 1 and 2? What would be the rule for p predictors $\{X_1, X_2, ...X_p\}$?

4. (1 Point) Recall from the lecture that for axis-aligned splits, there were a total of $p.(n-1)$ choices of splits in total. Where p is the number of dimensions and n are the number of samples. How many choices of splits do we have for the cases shown in parts 1 and 2 with n samples and $p = 2$?

5. (3 Points) Using the information from the lecture as well as the previous parts of this question, write down a pseudocode to learn classification trees with non axis-aligned splits. Make sure to specify which loss function you would use and justify why. You do not need to include the pruning step.

6. (1 Point) How many choices of splits are there in total for n samples and $p \in \{3, 5, 100\}$ with $n >> p$?

7. (0.5 x 2 Points) Using the answer from the previous part, give two reasons why the proposed approach may not work.

8. (0.5 x 2 Points) For each of the two reasons specified above, provide a solution that we could use to make the suggested method feasible.