15.095: Machine Learning Under a Modern Optimization Lens **Fall 2025**

Homework 3: Due October 20

Hand in: **pdf** upload to Gradescope. Please append any Julia code **at the end** of the whole pdf.

Note: this homework covers the Lecture 8-12 and Recitations 5-6. We recommend attempting questions after the relevant content has been covered in class.

Question 1: True/False (30 points)

Please classify the following statements as true or false and justify your answer. If the statement is false, please provide a counter example. We will assign 2 points for correctly classifying the answer, and 4 points for the validity of the justification/counterexample (for example, giving the correct reformulation).

(a) (6 Points) Suppose the data $\mathbf{X}$ are subject to perturbation $\mathbf{\Delta}$ and we want to add robustness to our features. Consider a row-wise uncertainty set where $\|\mathbf{\Delta}_i\|_2 \leq \rho$ for all observation i. The OCT (with parallel splits) accounting for such perturbation can still be formulated as a mixed-integer linear optimization problem with no additional constraints or variables.

(b) (6 points) Suppose we are given a dataset with severe class imbalance and we want to penalize misclassifications of class A twice as much as misclassifications of class B. In order to capture this in an OCT, we would need additional linear constraints to the original mixed-integer formulation.

(c) (6 points) Optimal classification trees with hyperplanes (OCT-Hs) with sparsity one are equivalent to optimal classification trees (OCTs).

(d) (6 points) Consider a prescriptive problem of

$$\min_{z \in \mathcal{Z}} \mathbb{E}[c(z; y)|x]. \tag{3.1}$$

Suppose $c(z; y) = 2zy$ and $y|x$ is uniformly distributed on $[0, x]$. Then, the prescriptive problem is equivalent to

$$\min_{z \in \mathcal{Z}} c(z; x). \tag{3.2}$$

Hint: You may use the following fact:

$$X \sim \text{Unif}(a, b) \implies \mathbb{E}[X] = \frac{b + a}{2}.$$

(e) (6 points) Consider the prescriptive problem

$$\max_{z \in \{1,-1\}} \sum_{i=1}^{n} w(x_i, x) c(y_i; z). \tag{3.3}$$

Suppose $c(y_i; z) = y_i z$. Then, using the KNN weighting method with k neighbors, the optimal solution to the prescriptive problem above is equivalent to setting

$$z = \text{sign}\Big(\sum_{i: x_i \in \text{KNN}(x)} y_i\Big). \tag{3.4}$$

Question 2: Extensions of the Greedy CART Algorithm (30 Points)

In this problem, you will greedily fit classification trees (using CART) to the data distribution given in Figure 3.1 (whereby $n = 28$ and $p = 2$); then, you will use the fitted tress for prescriptions.

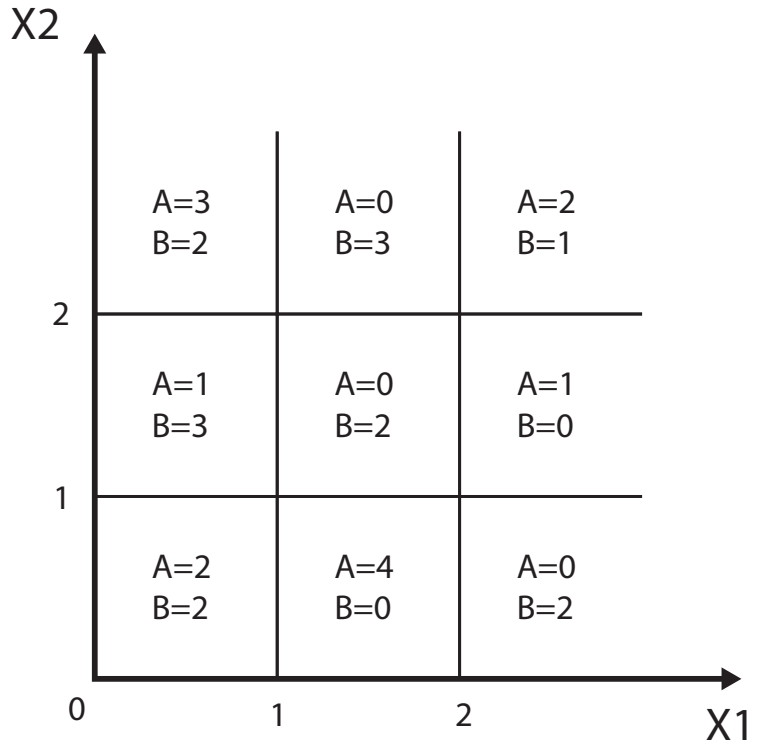

Figure 3.1: Distribution of data of two classes (A and B) in each of the regions.

For interpretability, your tree will need to satisfy the following requirements:

1. **Maximum depth:** Restrict the tree depth to 2.
2. **Integer splits:** Splits can only be performed at integer points.
3. **Variable ordering requirement:** You are given sets of features $\mathcal{F}_1, \ldots, \mathcal{F}_\kappa,\ 1 < \kappa < D$, such that $\mathcal{F}_k \cap \mathcal{F}_l = \emptyset$ and $\cup_{l=1}^{\kappa} \mathcal{F}_l = \{1, \ldots, p\}$. All features from set $\mathcal{F}_k$ are not allowed to be selected at split nodes unless, for all $l < k$, there exists some feature $i \in F_l$ which has already been selected in a preceeding split node in the same path. In our simple, 2-dimensional example, let $\mathcal{F}_1 = \{1\}, \mathcal{F}_2 = \{2\}$. You need to satisfy this requirement in building your CART tree.
4. **Cost Tie-breaker**: If multiple splits minimize the target error (missclassification or gini), then you should choose the split with the lowest cutoff value (e.g. a split at $X_1 = 1$ gains priority over a split at $X_1 = 2$ if those have the same error).

(a) (6 Points) Greedily build a tree using the misclassification error at each split while satisfying the requirements.

(b) (6 Points) Greedily build a tree using the gini impurity at each split while satisfying the requirements. Recall that the formula to compute the Gini impurity for a node t is $I_t = 1 - \sum_i^J p_j^2$, where p_j is the probability of finding label j in node t.

(c) (5 Points) Compare the resulting trees. Which has a better misclassification error? What does this tell us about using miscalssification vs gini impurity to build trees? Which is a better metric and why?

(d) (13 Points) Finally, you will use your trained **gini CART tree** to make prescriptions using the weighted-cost prediction/prescription formulation we saw in class. Consider the binary variable $z_i \in \{0, 1\}$ indicating whether subject (i.e., data point) i received a specific treatment or not. Depending on a subject's class, the cost of giving them the treatment or not is as follows:

c(z;y)	*y=A*	*y=B*
z=0	0	4
z=1	9	2

Consider a situation in which 9 new subjects, one from each cell/region of Figure 3.1 (recall that each subject's region is decided based on their features), are candidates to receive the treatment. Specifically, the i-th subject is mapped to the i-th region of Figure 3.2 (i.e. subject 1 lies in region R1, subject 2 in region R2 etc.).

However, you only have 5 treatments available.

Develop an integer linear optimization formulation to decide who will get the treatment. Solve (by hand) this formulation using the trained gini tree. Who gets the treatment?

Hint: Consider writing the cost functions $c(z; y = A)$ and $c(z; y = B)$ as linear functions of z.

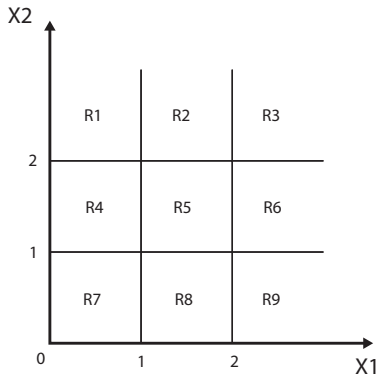

Figure 3.2: Region ID that corresponds to each cell

Question 3: Tree-based models (40 Points)

For this question, we are working with the CDC Diabetes Health Indicators Dataset. This dataset is derived from the 2005 Behavioral Risk Factor Surveillance System (BRFSS), an annual health-related telephone survey conducted by the CDC. We will apply tree-based methods to examine whether survey responses can predict an individual's diabetes status.

The dataset contains a variety of demographic and health-related features collected through the BRFSS survey. The target variable is `Diabetes_binary`, which indicates whether an individual has been diagnosed with diabetes (1) or not (0). The predictor variables include:

- `HighBP`: Indicator for high blood pressure.
- `HighChol`: Indicator for high cholesterol.
- `CholCheck`: Whether the individual has had their cholesterol checked in the past five years.
- `BMI`: Body Mass Index, a measure of weight relative to height.
- `Smoker`: Whether the individual has smoked at least 100 cigarettes in their lifetime.
- `Stroke`: Whether the individual has ever experienced a stroke.
- `HeartDiseaseorAttack`: History of coronary heart disease or myocardial infarction.
- `PhysActivity`: Participation in physical activity or exercise in the past 30 days.
- `Fruits`: Whether the individual consumes fruits at least once per day.
- `Veggies`: Whether the individual consumes vegetables at least once per day.
- `HvyAlcoholConsump`: Indicator for heavy alcohol consumption (more than 14 drinks per week for men and more than 7 for women).
- `AnyHealthcare`: Whether the individual has any form of health care coverage.
- `NoDocbcCost`: Whether the individual was unable to see a doctor in the past year due to cost.
- `GenHlth`: Self-rated general health status on a scale from 1 (excellent) to 5 (poor).
- `MentHlth`: Number of days in the past 30 days when mental health was not good.
- `PhysHlth`: Number of days in the past 30 days when physical health was not good.
- `DiffWalk`: Whether the individual has serious difficulty walking or climbing stairs.
- `Sex`: Biological sex (0 = female, 1 = male).
- `Age`: Categorical age variable (grouped into ranges, e.g., 18–24, 25–29, etc.).
- `Education`: Highest level of education attained (coded from 1 = never attended school to 6 = college graduate).
- `Income`: Income category (coded from 1 = less than \$10,000 to 8 = \$75,000 or more).

We will start by investigating how different choices of parameters for Optimal Classification Trees (OCTs) can influence the final tree model. Feel free to explore the `IAI` software documentation here. Please make sure to set the seed random seed = 15095 for all models and data splits.

(a) (5 Points) Train OCTs with depth=3 and minbucket=5, using each of the loss functions (misclassification, gini, and entropy). Discuss the differences in the resulting trees and report the AUC performance on the validation set.

(b) (10 Points) Train OCTs using the gini loss function and minbucket=5, for varying depths in [2, 4, 6, 8]. Plot the AUC validation scores as a function of the tree depth. Plot the complexity parameter selected by the algorithm as a function of the tree depth (note: when you use `GridSearch` from `IAI` autotunes the cp parameter for you. You just have to extract it from the final model). Discuss the results. Does the cp parameter tend to increase or decrease as the depth gets larger?

(c) (10 Points) Similarly, fix the depth and loss function, and plot the validation scores as a function of varying minbuckets in [5, 10, 20, 30]. As before, discuss the results.

(d) (25 Points) Now we investigate the dataset further and find the machine learning model with the best out-of-sample performance. You should appropriately validate hyperparameters and use a robust procedure for fairly deciding which model performs best. Implement and tune the following methods (any package is welcome; `IAI` implements all of them):

- CART (Classification and Regression Trees)
- Optimal Classification Trees (OCT)
- Random Forest
- Boosted Trees (e.g., XGBoost)
- Sparse Logistic Regression (Optimal Feature Selection Classifier in `IAI`)

When relevant, search the following hyperparameter grids:

$$\text{max\_depth} \in \{2, 4, 6, 8\}, \quad \text{minbucket} \in \{5, 10, 20, 30\}$$

$$\text{num\_trees} \in \{20, 25, 50, 100\}, \quad \text{num\_estimators} \in \{20, 25, 50, 100\}. \quad \text{sparsity} \in \{10, 15, 20\}.$$

Overall rubric

i. (5 points) Exzplain your methodology to tune the hyperparameters of your models.

ii. (10 points) Report the out-of-sample AUC and accuracy of each model on the test set.

iii. (10 points) Comment on interpretability/explainability of each model and discuss which variables were used by the different models. Which model would you choose as the best one?