15.095: Machine Learning Under a Modern Optimization Lens **Fall 2021**

Homework 6: Due November 24

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

6.1 Question 1: Interpretable Clustering (40 Points)

In this problem, we utilize the **Heart Failure Clinical Records Data Set**, a dataset that contains the medical records of 299 patients who had heart failure, collected during their follow-up period.

The original dataset, which you will find in the file `heart_failure.csv`, consists of about 300 data points and the following 13 features:

- Age: age of the patient (years).
- Anaemia: decrease of red blood cells or hemoglobin (boolean).
- High blood pressure: if the patient has hypertension (boolean).
- Creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L).
- Diabetes: if the patient has diabetes (boolean).
- Ejection fraction: percentage of blood leaving the heart at each contraction (percentage).
- Platelets: platelets in the blood (kiloplatelets/mL).
- Sex: woman or man (binary).
- Serum creatinine: level of serum creatinine in the blood (mg/dL).
- Serum sodium: level of serum sodium in the blood (mEq/L).
- Smoking: if the patient smokes or not (boolean).
- Time: follow-up period (days).
- Death event: if the patient deceased during the follow-up period (boolean).

To better understand the factors that are critical for different types of heart failure patients, we are interested in clustering our data. Contrary to the unified ICOT (Interpretable Clustering via Optimal Trees) approach we saw in class, here we will apply a two-stage approach to interpret our clusters.

(a) (5 Points) After loading the data, calculate basic statistics (maximum, minimum, mean, median, standard deviation) for each feature. Explain how you need to preprocess the data prior to performing clustering and apply this preprocessing.

(b) (15 Points) Cluster the data using your preferred clustering algorithm (e.g., k-means). Describe your approach to select the number of clusters.

Note: Feel free to use any package of your choice. You can look into the `Clustering.jl` *package, which is part of the* `JuliaStats` *collection.*

(c) (5 Points) Create a plot of your choice to visualize the clustering (e.g., plot a feature A with respect to a feature B, coloring by cluster).

(d) (5 Points) Train an Optimal Classification Tree mapping data points to their assigned clusters. Explain your methodology to choose the hyperparameters of your tree.

(e) (5 Points) Use the learned tree to provide a description for each cluster.

(f) (5 Points) Discuss any problems that you observe in the two-stage approach, which you could have avoided by using ICOT.

6.2 Question 2: The Power of Optimization Over Randomization (60 marks)

In this question, let us investigate how an MIO approach compares with other random allocation methods in assigning subjects to different groups to minimize discrepancies. Suppose we would like to split the top 30 Men's singles tennis players into groups during the first round such that each group has a similar distribution of player levels, so all players have a fair chance of going into the second round. We use their ATP points [1] as an estimation for their abilities, found in players.csv. Please standardize the data following the steps in Lecture 14, specifically:

$$w_i' = (y_i - \hat{\mu})/\hat{\sigma}, \quad \text{where} \quad \hat{\mu} = \sum_{i=1}^{n} y_i/n \quad \text{and} \quad \hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{\mu})^2/n.$$

(a) Suppose we want to split the numbers into three groups (with 10 players in each group) to minimize the discrepancies in centered first and second moments as we have learned in class. Please implement the following algorithms 1 to 4. For each algorithm, report the discrepancies in centered first and second moment, that is, for each pair of groups p, q, report $|\mu_p(x) - \mu_q(x)|$, $\left|\sigma_p^2(x) - \sigma_q^2(x)\right|$, as well as the maximum and mean discrepancy in each moment. **Discuss the results.**

(i) **(3 marks)** Randomization: Shuffle all numbers. Split the group based on shuffled index, i.e., first third goes to the first group, second third to the second group, etc.

(ii) **(3 marks)** Re-randomization: Repeat Algorithm 1 with $10,000$ different seeds, and choose the group split with the lowest sum of the absolute difference of the first and second moments in the three groups.

(iii) **(4 marks)** Triplet Matching: Rank all players by increasing order of matches won. For every consecutive three players, randomly split them into the first, second, or third group.

(iv) **(20 marks)** Optimization: Solve this mixed-integer optimization problem using the formulation we have seen from the lecture notes. You can set $\rho = 0.5$.

(b) Suppose we want to split the data into three groups (with 10 players in each group) to minimize the sum of pairwise difference between the player levels in each group. In this case, our objective function is

$$\min \sum_{p \neq q} \sum_{i \in \text{ group p},\ j \in \text{ group q}} \left|w_i' - w_j'\right| \tag{6.1}$$

[1] `https://www.atptour.com/en/rankings/singles?rankDate=2021-11-15&countryCode=all&rankRange=0-100&sort=points&sortAscending=False`

(i) **(10 marks)** Please formulate (6.1) as a mixed-integer linear optimization problem. Make sure to write down and **explain** your mixed-integer linear formulation in your hand-in (Julia code is not sufficient).

(ii) **(20 marks)** Implement your MIO approach in Julia, and report the value of the objective function in (6.1) for this optimization approach and also for the Re-Randomization approach 2 in (a). **Discuss your findings.**