Math 126 Problem 22: Leverage Scores of Shapley Value Regression

Spring 2027

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Monday, May 3, 2027 (Pacific time).**

- You are encouraged to discuss ideas and work with your classmates. However, you *must acknowledge* your collaborators at the top of each solution on which you collaborated with others and you *must write* your solutions independently.
- Your solutions to theory questions must be written legibly, or typeset in LaTeX or markdown.
- I recommend that you write your solutions in a Python notebook using Google Colab.
- You should submit your solutions as a *single PDF* via the assignment on Gradescope.

Grading: The point of the problem set is for *you* to learn. To this end, I hope to disincentivize the use of LLMs by *not* grading your work for correctness. Instead, you will grade your own work by comparing it to my solutions. This self-grade is due the Friday *after* the problem set is due, also on Gradescope.
Original notebook: Google Colab.

Leverage Scores of Shapley Value Regression

Shapley values, originally from cooperative game theory and now widely used in explainable AI, can be uniquely defined as the optimal solution to a weighted linear regression problem.

Let $v : 2^{[d]} \to \mathbb{R}$ be a value function mapping subsets of d features to a real number, with $v(\emptyset) = 0$. The vector of Shapley values $\boldsymbol{\phi} \in \mathbb{R}^d$ minimizes the weighted least squares objective:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d : \langle \boldsymbol{\beta}, \mathbf{1} \rangle = v([d])} \sum_{S \subseteq [d] : 0 < |S| < d} \left(v(S) - \sum_{i \in S} \beta_i \right)^2 w_S$$

where the weights are $w_S = \frac{1}{\binom{d}{|S|} |S|(d-|S|)}$.

To analyze this system, we can convert it into an unconstrained ordinary least squares problem. Define the uncentered indicator matrix $\mathbf{A}' \in \mathbb{R}^{(2^d-2) \times d}$ such that its (S, i)-th entry is $\mathbb{1}[i \in S]\sqrt{w_S}$. We also define the centering projection matrix $\mathbf{P} = \mathbf{I} - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$. The centered design matrix for our regression is given by:

$$\mathbf{A} = \mathbf{A}'\mathbf{P}$$

In this problem, you will analyze the geometry of this regression by computing the **leverage scores** of the matrix $\mathbf{A}$. The leverage score ℓ_S for a row $\mathbf{a}_S^\top$ is defined as the S-th diagonal element of the hat matrix:

$$\ell_S = \mathbf{a}_S^\top (\mathbf{A}^\top \mathbf{A})^+ \mathbf{a}_S$$

where $(\mathbf{A}^\top \mathbf{A})^+$ is the Moore-Penrose pseudoinverse.

Part A: The Uncentered Covariance Matrix

Let $\mathbf{M} = \mathbf{A'}^\top \mathbf{A'}$. Argue that $\mathbf{M}$ is permutation invariant, meaning its diagonal entries are all equal to some constant α, and its off-diagonal entries are all equal to some constant β.

Then, express $\mathbf{M}$ in the form $c\mathbf{I} + \gamma \mathbf{1}\mathbf{1}^\top$ in terms of α and β.

Part B: The Centered Covariance Matrix

Using the fact that $\mathbf{A} = \mathbf{A'P}$ and your result from Part A, show that the centered covariance matrix simplifies to $\mathbf{A}^\top \mathbf{A} = c\mathbf{P}$, where $c = \alpha - \beta$. *(Recall that* $\mathbf{P1} = \mathbf{0}$ *and* $\mathbf{P}^2 = \mathbf{P}$*.)*

Next, evaluate c by computing the quadratic form $\mathbf{v}^\top (\mathbf{A}^\top \mathbf{A})\mathbf{v}$ for the difference vector $\mathbf{v} = \mathbf{e}_i - \mathbf{e}_j$ (where $i \neq j$).

Hint: Evaluate $\|\mathbf{A'v}\|^2$ *by summing over the rows of* $\mathbf{A'}$*, and use the combinatorial identity* $\frac{\binom{d-2}{k-1}}{\binom{d}{k}} = \frac{k(d-k)}{d(d-1)}$ *to show that* $c = \frac{1}{d}$*.*

Part C: The Pseudoinverse

Recall that $\mathbf{P}$ is a projection matrix. Given that $\mathbf{A}^\top \mathbf{A} = \frac{1}{d}\mathbf{P}$, what is the Moore-Penrose pseudoinverse $(\mathbf{A}^\top \mathbf{A})^+$? Briefly justify your answer.

Part D: Computing the Leverage Scores

The S-th row of $\mathbf{A}$ is given by $\mathbf{a}_S^\top = \sqrt{w_S}\mathbf{1}_S^\top \mathbf{P}$, where $\mathbf{1}_S \in \{0,1\}^d$ is the binary indicator vector for the subset S. Substitute $\mathbf{a}_S$ and $(\mathbf{A}^\top \mathbf{A})^+$ into the leverage score formula $\ell_S = \mathbf{a}_S^\top (\mathbf{A}^\top \mathbf{A})^+ \mathbf{a}_S$.

Expand the resulting quadratic form $\mathbf{1}_S^\top \mathbf{P} \mathbf{1}_S$ and substitute the definition of the Shapley weights w_S to prove that:

$$\ell_S = \frac{1}{\binom{d}{|S|}}$$

Interpretation: What does this final result tell you about the total sum of leverage scores for all subsets of a fixed size k? How does the total sum of *all* leverage scores relate to the rank of the matrix $\mathbf{A}$?