Week 13

↳ Pset 12
- due Sunday (hard deadline)
- self-grade due Monday (hard deadline)

↳ Anyone interested in grading next semester?

↳ Midterm 11/25

Suggestion:
1. Review psets (including this week's)
2. Review recurring concepts (e.g., SVD, variance, etc)
3. Take 1st practice exam (75 min on paper). Review.
4. Take 2nd.

Explainable AI

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \longrightarrow \boxed{f} \rightarrow f(x)$$

Q: <u>why</u> did the model output $f(x)$?

Idea: Let's compare to a baseline $b \in \mathbb{R}^d$
(e.g., an average point)

Special Case: Linear Models

$$f(x) = \sum_{i=1}^{d} w_i x_i \quad \text{vs.} \quad f(b) = \sum_{i=1}^{d} w_i b_i$$

$$f(x) - f(b) = \sum_{i=1}^{d} w_i (x_i - b_i)$$

A: Feature i changed output by

$$\phi_i = w_i (x_i - b_i)$$

Motivation: Use ϕ_i to...

- Check for concerning behavior
- Understand model
- Select important features

General Case: Non-linear models

Challenge: Complicated interactions btwn features

E.g. Suppose f predicts plant growth.
If temp is high, precip helps.
If temp is low, precip hurts.

Solution: Evaluate effect of feature i in context of other features

Value Function

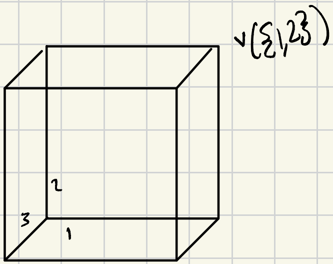
$S \subseteq \{1, \dots, d\}$

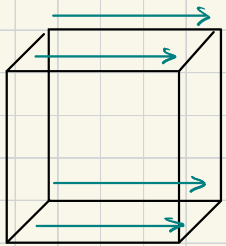
$x^S \in \mathbb{R}^d$ combines x, b

$$x_i^S = \begin{cases} x_i & \text{if } i \in S \\ b_i & \text{if } i \notin S \end{cases}$$

$v(S) = f(x^S)$

Goal: Analyze $\mathbb{E}_S\left[v(S \cup i) - v(S)\right]$

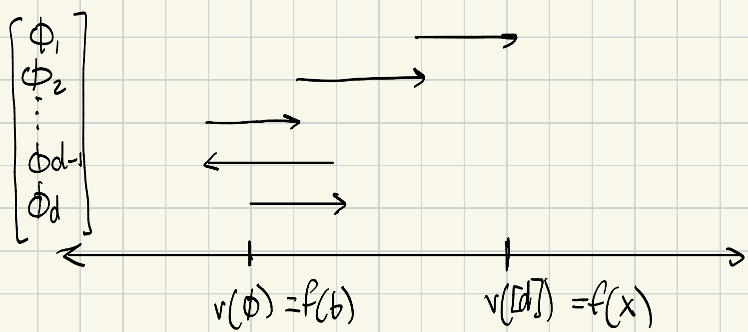




Shapley Values

Axioms:
- ↳ Null
- ↳ Symmetry
- ↳ Linearity
- ↳ Efficiency: $\sum_{i=1}^{d} \phi_i = v([d]) - v(\emptyset)$

$$\phi_i = \sum_{S \subseteq [d] \setminus \{i\}} \left[v(S \cup i) - v(S)\right] \frac{1}{d\binom{d-1}{|S|}}$$

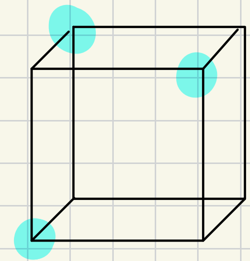


Decompose change in prediction in terms of features

Estimating Shapley Values

Challenge: 2^d subsets to evaluate

Monte carlo Estimator approximates sum by computing several terms



Sample m subsets instead of all 2^d

$S_1, \dots, S_m \subseteq [d] \setminus \{i\}$

S is sampled w.p. p_S without replacement

Choose $p_S = \frac{1}{d\binom{d-1}{|S|}}$

$$\tilde{\phi}_i^{MC} = \frac{1}{m} \sum_{j=1}^{m} [v(S_j \cup i) - v(S_j)] \frac{1}{d\binom{d-1}{|S_j|} p_{S_j}} = \frac{1}{m} \sum_{j=1}^{m} v(S_j \cup i) - v(S_j)$$

Goal:

- unbiased
- low variance

$$\mathbb{E}[\tilde{\phi}_i^{MC}] = \mathbb{E}\Big[\frac{1}{m} \sum_{j=1}^{m} \sum_{S \subseteq [d] \setminus i} [v(S \cup i) - v(S)] \mathbb{1}[S_j = S] \Big]$$

$$= \dots$$

$$= \phi_i$$

Variance Bound

- Scalar $\quad Var(cX) = c^2 Var(X)$

Fact : Variance without replacement ≤ variance with replacement

So, $Var(\phi_i^{MC})$ at most variance when sampled with replacement

$$Var(\phi_i^{MC}) = Var \frac{1}{m} \sum_{j=1}^{m} \sum_{S \subseteq [d] \setminus i} [v(S \cup i) - v(S)] \, \mathbb{1}[S_j = S]$$

w/ replacement

$$= \frac{1}{m^2} \sum_{j=1}^{m} \sum_{S \subseteq [d] \setminus i} [v(S \cup i) - v(S)]^2 \, Var(\mathbb{1}[S_j = S]) \quad \text{by indep}$$

$$\leq \frac{1}{m} \sum_{S \subseteq [d] \setminus i} [v(S \cup i) - v(S)]^2 \, P_{|S|} \qquad \text{by below inequality}$$

Chebyshev's

$$Pr(|\phi_i^{MC} - \phi_i| \geq \epsilon) \leq \frac{Var(\phi_i^{MC})}{\epsilon^2}$$

$$Var(\mathbb{1}[S_j = S]) = \mathbb{E}[\mathbb{1}[S_j = S]^2] - \mathbb{E}[\mathbb{1}[S_j = S]]^2$$

$$\leq \mathbb{E}[\mathbb{1}[S_j = S]^2]$$

$$= 1 \cdot Pr(S_j = S) + 0 [1 - Pr(S_j = S)]$$

$$= Pr(S_j = S) = P_{|S|}$$

Maximum Sample Reuse

All this for only one $\hat{\phi}_i^{MC}$? Let's reuse samples!

$$\phi_i = \sum_{S \subseteq [d] \setminus \{i\}} [v(S \cup i) - v(S)] \frac{1}{d\binom{d-1}{|S|}}$$

$$= \sum_{S \subseteq [d] : i \in S} v(S) \frac{1}{d\binom{d-1}{|S|-1}} - \sum_{S \subseteq [d] : i \notin S} v(S) \frac{1}{d\binom{d-1}{|S|}}$$

$$= \sum_{S \subseteq [d]} v(S) \left[\frac{\mathbb{1}[i \in S]}{d\binom{d-1}{|S|-1}} - \frac{\mathbb{1}[i \notin S]}{d\binom{d-1}{|S|}} \right]$$

Estimate this! But variance will depend on $[v(S)]^2$ rather than $[v(S \cup i) - v(S)]^2$...

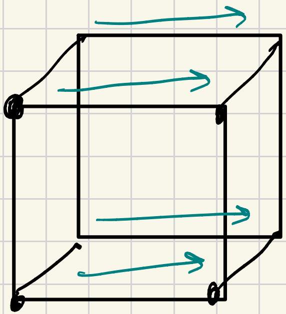
Last Lecture

↳ Midterm 11/25

↳ Important class on 12/2

- proposal due 10am
- discuss proposal in class
- class eval

Shapley Values



$v(S)$ ← • prediction given features in S

$$\phi_i = \sum_{S \subseteq [d]\setminus\{i\}} \frac{v(S \cup i) - v(S)}{d\binom{d-1}{|S|}}$$

= average effect of feature i on prediction

Challenge: 2^d subsets!!

Monte Carlo

Sample $S_1, \dots, S_m \subseteq [d]\setminus\{i\}$ wp $P_{|S|} = \frac{1}{d\binom{d-1}{|S|}}$

$$\hat{\phi}_i^{MC} = \frac{1}{m} \sum_{S \subseteq [d]\setminus\{i\}} \frac{v(S \cup i) - v(S)}{d\binom{d-1}{|S|} P_{|S|}} \mathbb{1}[S \text{ sampled}]$$

$$= \frac{1}{m} \sum_{S \subseteq [d]\setminus\{i\}} \big(v(S \cup i) - v(S)\big) \mathbb{1}[S \text{ sampled}]$$

Goal: Let's compute variance!

Variance

$$\mathrm{Var}(X+Y) = \mathbb{E}[(X+Y-\mathbb{E}[X]-\mathbb{E}[Y])^2]$$

$$= \underbrace{\mathbb{E}[(X-\mathbb{E}[X])^2]}_{\mathrm{Var}(X)} + \underbrace{\mathbb{E}[(Y-\mathbb{E}[Y])^2]}_{\mathrm{Var}(Y)} + 2\underbrace{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}_{\mathrm{Cov}(X,Y)}$$

If X, Y independent...

Variance of Monte Carlo

$$\hat{\phi}_i^{MC} = \frac{1}{m} \sum_{S \subseteq [d] \setminus \{i\}} \left(v(S \cup i) - v(S) \right) \mathbb{1}[S \text{ sampled}]$$

$$= \frac{1}{m} \sum_{S} x_S \mathbb{1}_S$$

$$\mathbb{E}[\mathbb{1}_S^2] - \mathbb{E}[\mathbb{1}_S]^2 \leq \mathbb{E}[\mathbb{1}_S^2]$$

$$= \mathbb{E}[\mathbb{1}_S]$$

$$= 1 \cdot Pr(S) + 0 \cdot (1 - Pr(S))$$

$$= Pr(S) = P_{|S|}$$

$$Var(\hat{\phi}_i^{MC}) = \frac{1}{m^2} \sum_S Var(x_S \mathbb{1}_S) + \frac{1}{m^2} 2 \sum_{S,S'} cov(x_S \mathbb{1}_S, x_{S'} \mathbb{1}_{S'})$$

$$= \frac{1}{m^2} \sum_S \mathbb{E}\left[(x_S \mathbb{1}_S - \mathbb{E}[x_S \mathbb{1}_S])^2\right] + \frac{2}{m^2} \sum_{S,S'} \mathbb{E}\left[(x_S \mathbb{1}[S] - \mathbb{E}\, x_S \mathbb{1}_S)(x_{S'} \mathbb{1}_{S'} - \mathbb{E}\, x_{S'} \mathbb{1}_{S'})\right]$$

$$= \frac{1}{m^2} \sum_S x_S^2 \left(\mathbb{E}[\mathbb{1}_S^2] - \mathbb{E}[\mathbb{1}_S]^2\right) + \frac{2}{m^2} \sum_{S,S'} x_S x_{S'} \mathbb{E}\left[(\mathbb{1}_S - \mathbb{E}[\mathbb{1}_S])(\mathbb{1}_{S'} - \mathbb{E}[\mathbb{1}_{S'}])\right]$$

$$\mathbb{E}\left[(\mathbb{1}_S - \mathbb{E}[\mathbb{1}_S])(\mathbb{1}_{S'} - \mathbb{E}[\mathbb{1}_{S'}])\right] = \mathbb{E}[\mathbb{1}_S \mathbb{1}_{S'}] - 2\,\mathbb{E}[\mathbb{1}_S]\,\mathbb{E}[\mathbb{1}_{S'}] + \mathbb{E}[\mathbb{1}_S]\,\mathbb{E}[\mathbb{1}_{S'}]$$

$$= Pr(\text{both } S, S' \text{ selected}) - Pr(S \text{ selected})\, Pr(S' \text{ selected}) \leq 0$$

$$Var(\hat{\phi}_i^{MC}) \leq \frac{1}{m^2} \sum_{S \subseteq [d] \setminus \{i\}} [v(S \cup i) - v(S)]^2 \, P_{|S|}$$

Maximum Sample Reuse

Motivation: We compute $v(S \cup i) - v(S)$,
but only use it to estimate ϕ_i

Unbiased? ✓ :)

Variance? depends on $[v(S)]^2$:(

$$\phi_i = \sum_{S \subseteq [d] \setminus \{i\}} \frac{v(S \cup i) - v(S)}{d\binom{d-1}{|S|}}$$

$$= \sum_{S : i \in S} \frac{v(S)}{d\binom{d-1}{|S|-1}} - \sum_{S : i \notin S} \frac{v(S)}{d\binom{d-1}{|S|}}$$

$$= \sum_{S} v(S) \left[\frac{\mathbb{1}[i \in S]}{d\binom{d-1}{|S|-1}} - \frac{\mathbb{1}[i \notin S]}{d\binom{d-1}{|S|}} \right]$$

Sample $S_1, \ldots, S_m \subseteq [d]$, use <u>for all</u> estimates

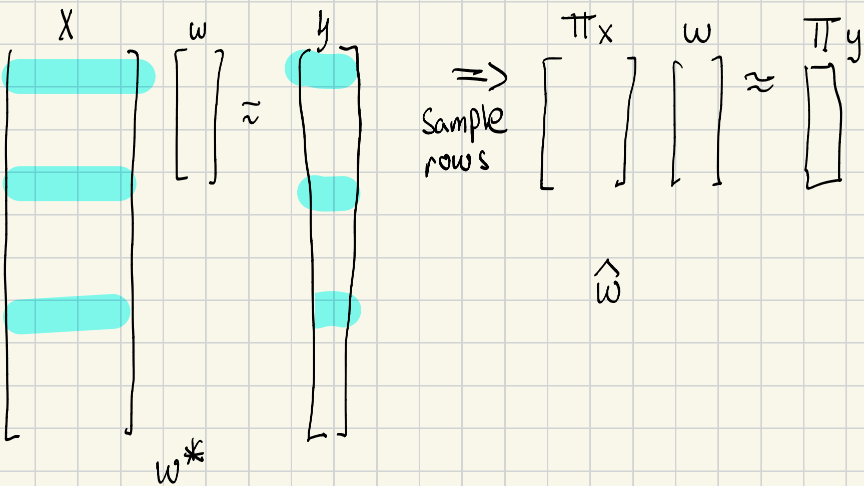
$$\hat{\phi}_i^{MSR} = \frac{1}{m} \sum_{S} v(S) \left[\frac{\mathbb{1}[i \in S]}{d\binom{d-1}{|S|-1}} - \frac{\mathbb{1}[i \notin S]}{d\binom{d-1}{|S|}} \right] \frac{\mathbb{1}[S \text{ sampled}]}{P_{[S]}}$$

Active Learning

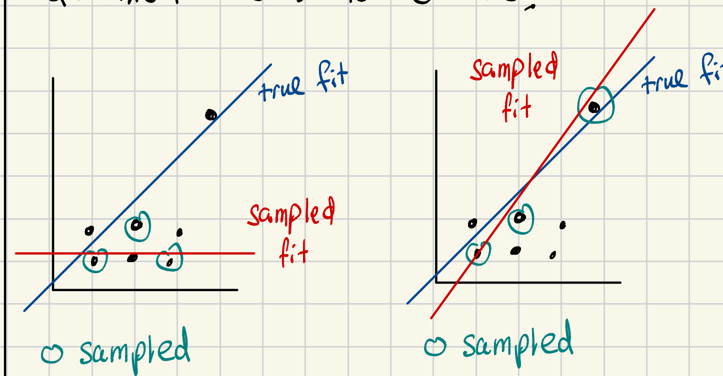
$X \in \mathbb{R}^{2^d \times d}$ $\qquad y \in \mathbb{R}^d$

$$\phi = \begin{bmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_d\end{bmatrix} = \underset{w:\langle w,1\rangle = v([d])-v(\emptyset)}{\arg\min} \|Xw-y\|_2^2$$

$$= \underset{w:\langle w,1\rangle = v(d)-v(\emptyset)}{\arg\min} \sum_S \Big(v(S) - \sum_{i:i\in S} w_i\Big)^2$$



Q: Which rows to choose?



Idea: Choose row by its leverage

If we sample $m = \tilde{O}\left(\frac{d}{\epsilon^2}\right)$ rows,

$$\|X\hat{w} - y\|_2^2 \le (1+\epsilon)\,\|Xw^* - y\|_2^2$$

Regression Adjustment

Sample $S_1, \dots, S_m$

Fit $\hat{v}$ to v on samples

Return $\phi_i(\hat{v}) + \hat{\phi}_i^{MSR}(v - \hat{v})$

Consistent: What is $\phi_i(\hat{v}) + \phi_i(v - \hat{v})$?