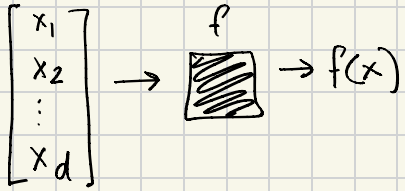
Week 13

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

↳ 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



Q: Why 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

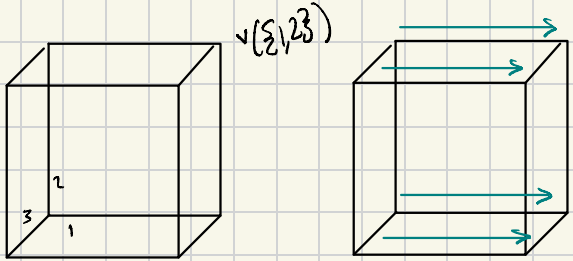
$S \subseteq \{1, \ldots, d\}$

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

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

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

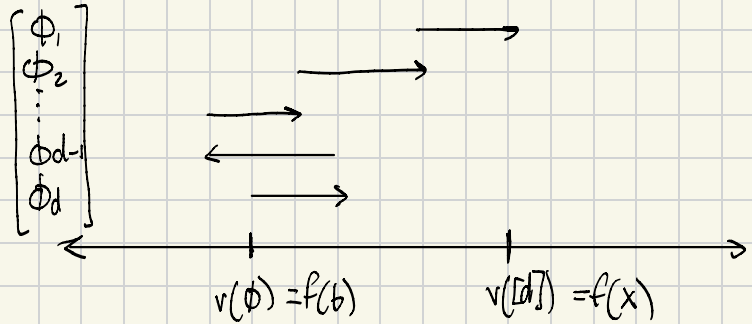
Goal: Analyze $\sum_S [v(S \cup i) - v(S)]$



Shapley Values

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

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

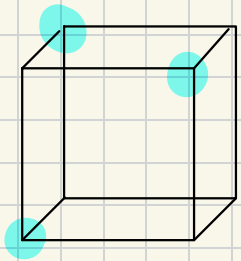


Decompose change in prediction in terms of features

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

$S_1, \dots, S_m \subseteq [S]$

S is sampled wp 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

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

$$= \dots$$

$$= \phi_i$$

Variance Bound

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

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

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

$$\leq \frac{1}{m}\sum_{S\subseteq[d]\setminus\{i\}}[v(S\cup i)-v(S)]^2\ p_S$$

Intuition: Variance when sampled without replacement $\leq$ varian when sampled with replacement

$$Var\sum_{j=1}^{m}\mathbb{1}[S_j = S] = \mathbb{E}\left[\left(\sum_{j=1}^{m}\mathbb{1}_j\right)^2\right] - \mathbb{E}\left[\sum_{j=1}^{m}\mathbb{1}_j\right]^2 \leq \mathbb{E}\left[\left(\sum_{j=1}^{m}\mathbb{1}_j\right)^2\right] = \mathbb{E}\sum_{j,k=1}^{m}\mathbb{1}_j\mathbb{1}_k$$

($\mathbb{1}_j$ = "$\mathbb{1}[S_j = S]$")

$$= \sum_{j=1}^{m}\mathbb{E}[\mathbb{1}_j^2] + \sum_{j\neq k}\mathbb{E}[\mathbb{1}_j\mathbb{1}_k]$$

↑ Pr(S selected for j and k) = 0

$$\leq \sum_{j=1}^{m} p_S = m\,p_S$$

Chebyshev's

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

Maximum Sample Reuse

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

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

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

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

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