

Week 13

↳ Pset 12

- due Sunday (hard deadline)
- self-grade due Monday (hard deadline)

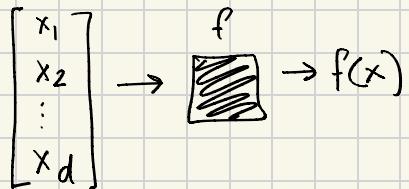
↳ Anyone interested in grading next semester?

↳ Midterm 11/25

Suggestion:

1. Review psets (including this week!)
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 b/wn 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

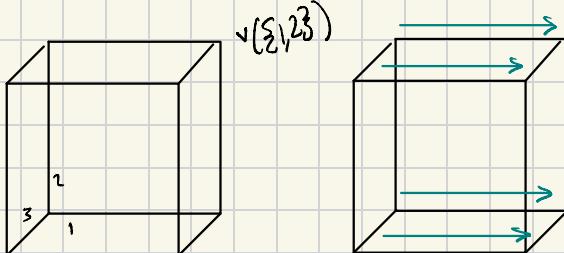
$$S \subseteq \{1, \dots, d\}$$

$x^S \in \mathbb{R}^d$ 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 $\sum_S [v(S \cup i) - v(S)]$



Shapley Values

Axioms:

↳ Null

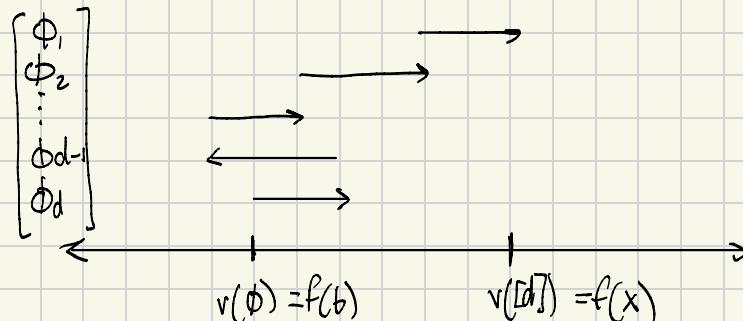
↳ Symmetry

↳ Linearity

↳ Efficiency

$$\sum_{i=1}^d \phi_i = v(\{d\}) - v(\emptyset)$$

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

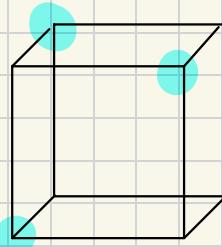


Decompose change in prediction in terms of features

Estimating Shapley Values

Challenge: 2^d subsets to evaluate

Monte Carlo Estimator approximates sum by computing several terms



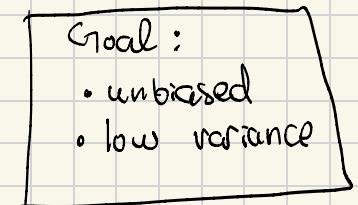
Sample m subsets instead of all 2^d

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

S is sampled w/o replacement

$$\text{Choose } P_S = \frac{1}{d \binom{d-1}{|S|-1}} \\ \tilde{\Phi}_i^{MC} = \frac{1}{m} \sum_{j=1}^m [v(S_j \cup i) - v(S_j)] \frac{1}{d \binom{d-1}{|S_j|-1} P_{S_j}} = \frac{1}{m} \sum_{j=1}^m v(S_j \cup i) - v(S_j)$$

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



$$= \Phi_i$$

Variance Bound

- Scalar $\text{Var}(cX) = c^2 \text{Var}(X)$

Fact: Variance without replacement \leq variance with replacement

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

$$\begin{aligned}\text{Var}(\phi_i^{MC}) &= \text{Var} \left(\frac{1}{m} \sum_{j=1}^m \sum_{S \subseteq [d] \setminus i} [\nu(s_{v(i)}) - \nu(S)] \mathbb{I}[S_j = S] \right) \\ &\stackrel{\uparrow \text{ w/ replacement}}{=} \frac{1}{m^2} \sum_{j=1}^m \sum_{S \subseteq [d] \setminus i} [\nu(s_{v(i)}) - \nu(S)]^2 \text{Var}(\mathbb{I}[S_j = S]) \quad \text{by indep} \\ &\leq \frac{1}{m} \sum_{S \subseteq [d] \setminus i} [\nu(s_{v(i)}) - \nu(S)]^2 p_{ISI} \quad \text{by below inequality}\end{aligned}$$

$$\begin{aligned}\text{Var}(\mathbb{I}[S_j = S]) &= \mathbb{E}[\mathbb{I}[S_j = S]^2] - \mathbb{E}[\mathbb{I}[S_j = S]]^2 \\ &\leq \mathbb{E}[\mathbb{I}[S_j = S]^2] \\ &= 1 \cdot \Pr(S_j = S) + O[1 - \Pr(S_j = S)] \\ &\approx \Pr(S_j = S) = p_{ISI}\end{aligned}$$

Chebyshev's

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

Maximum Sample Reuse

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

$$\begin{aligned}\hat{\phi}_i &= \sum_{S \subseteq [d] \setminus \{i\}} [v(S \cup i) - v(S)] \frac{1}{d \binom{d-1}{|S|}} \\ &= \sum_{S \subseteq [d]: i \in S} v(S) \frac{1}{d \binom{d-1}{|S|-1}} - \sum_{S \subseteq [d]: i \notin S} v(S) \frac{1}{d \binom{d-1}{|S|}} \\ &= \sum_{S \subseteq [d]} v(S) \left[\frac{\mathbb{I}[i \in S]}{d \binom{d-1}{|S|-1}} - \frac{\mathbb{I}[i \notin S]}{d \binom{d-1}{|S|}} \right]\end{aligned}$$

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