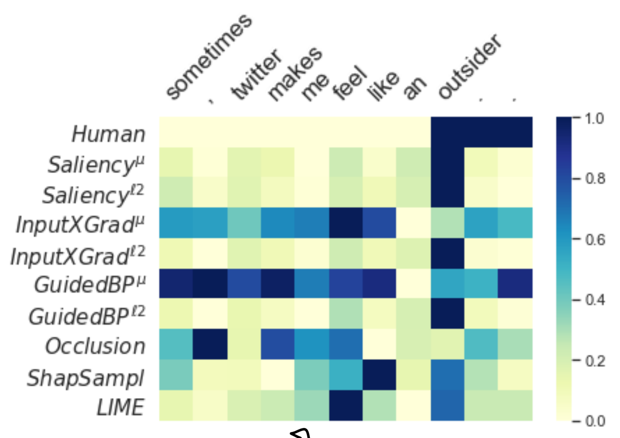
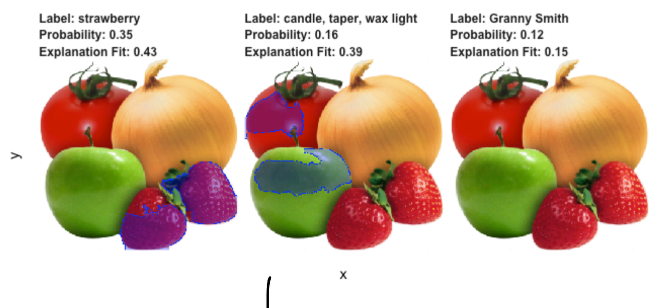
Week 3 - LIME & SHAP

Wednesday, February 5, 2025 12:06 PM

1. LIME and SHAP are both methods for generating local explanations. For a given sample, what are the most important features?



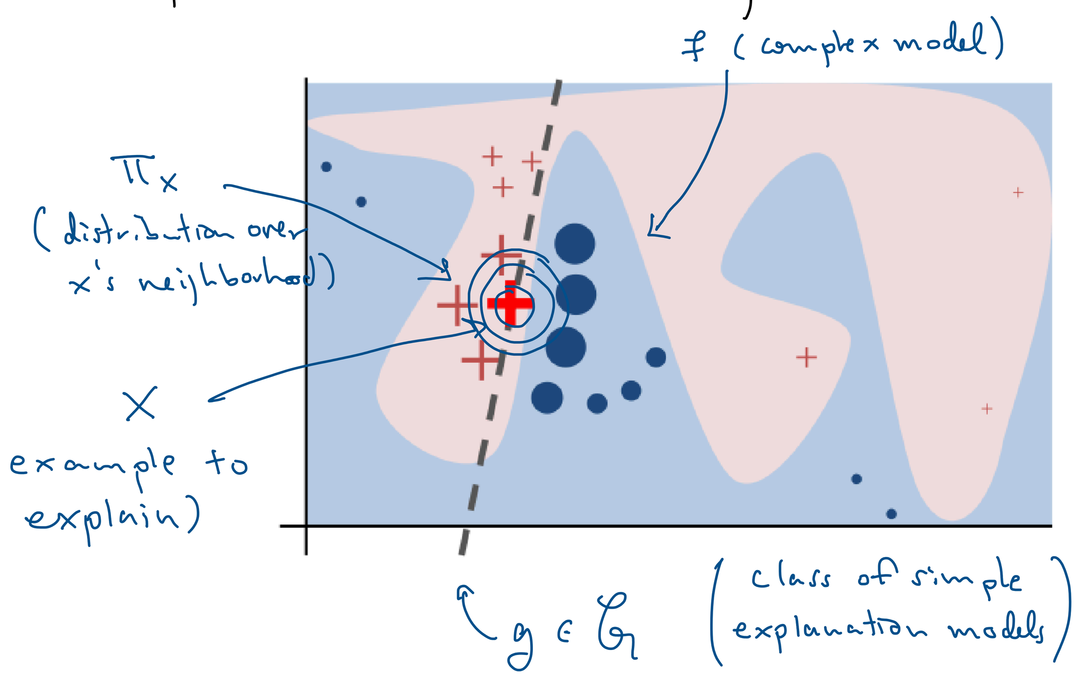
This sentence is predicted to be negative sentiment — "outsider" seems important.



"Candle" is the 2nd most probable class. The reflection might be the reason.

LIME

2. Intuition: Approximate complex boundaries with planes (centered at the sample we want to explain.



3. Once the forms of π_x and G are specified, follow:

(i) Sample $x_n' \sim \pi_x$

(ii) Summarize $x_n' \longrightarrow z_n'$

(iii) Solve

$$\min_{g \in G} \frac{1}{N}\sum_{n=1}^{N} L\left(f(z_n'), g(z_n')\right) + \Omega(g)$$

4. Example: For the sentiment prediction problem, we could use,

π_x - Select sentences with high cosine similarity to the target

$x_n \to z_n$ - Transform to word counts

g - linear model

Ω - ℓ^1 regularization

$$\Longrightarrow \frac{1}{N}\sum_{n=1}^{N}\left(f(z_n) - z_n^T\beta\right)^2 + \lambda\|\beta\|$$

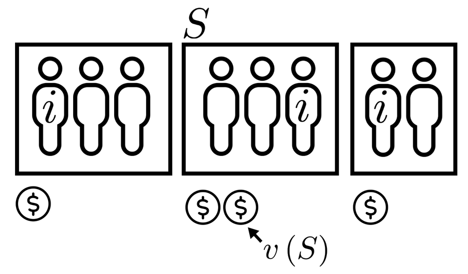
and β tells us which words affect sentiment locally at this example.

5. Challenges: So many hyperparameters!

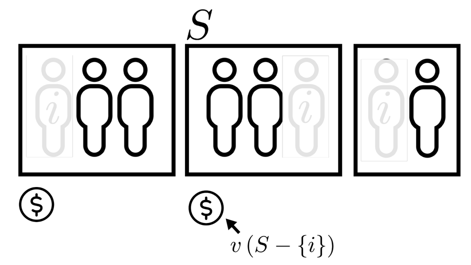
- Best summarizer $x_n \to z_n$?
- How to define a neighborhood? How large?
- Which g for which f?

SHAP - Conceptual

6. SHAP is motivated by the credit assignment problem.



Team S has profit $v(S)$. How much to give person i?



idea: How much does profit decrease when removing person i?

7. We formalize this as

$$\varphi(i) = \underbrace{\frac{1}{D}\sum_{d=1}^{D}}_{\text{Vary team sizes}} \overbrace{\frac{1}{\binom{D-1}{d-1}}\sum_{S \in S_d(i)}}^{\text{Avg. over } S_d(i)} \overbrace{\left[v(S) - v(S - \{i\})\right]}^{\text{Profit Decrease}}$$

$S_d(i) := \{\text{subsets of size } d \text{ including person } i\}$

8. What does this have to do w/ Local Explanation?

- employee $\longrightarrow$ feature
- Team $\longrightarrow$ Subset of features
- Profit $\longrightarrow$ Expected prediction

More formally, to explain f at sample x,

$$v_x(S) = \mathbb{E}_{p(x'_{S^c}|x_S)}\left[f(x_S, x_{S^c})\right]$$

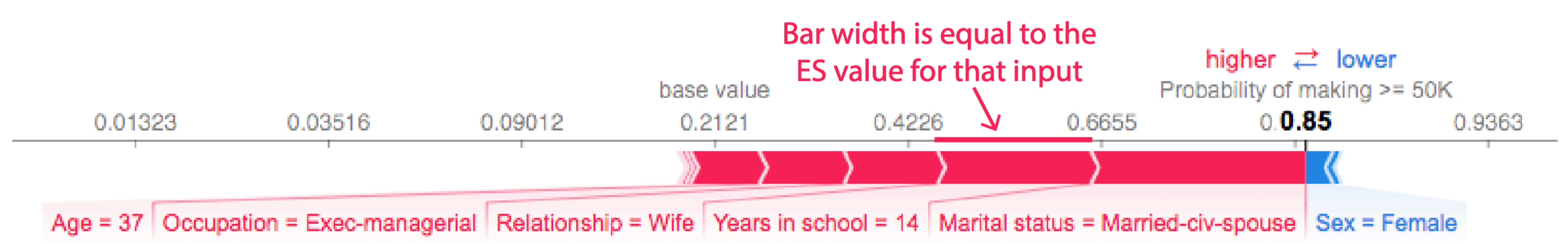
Fix x_S at current sample's values. Average over other features.

$$\varphi_x(f, i) = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{\binom{D-1}{d-1}}\sum_{S \in S_d(i)}\left[v_x(S) - v_x(S - \{i\})\right]$$

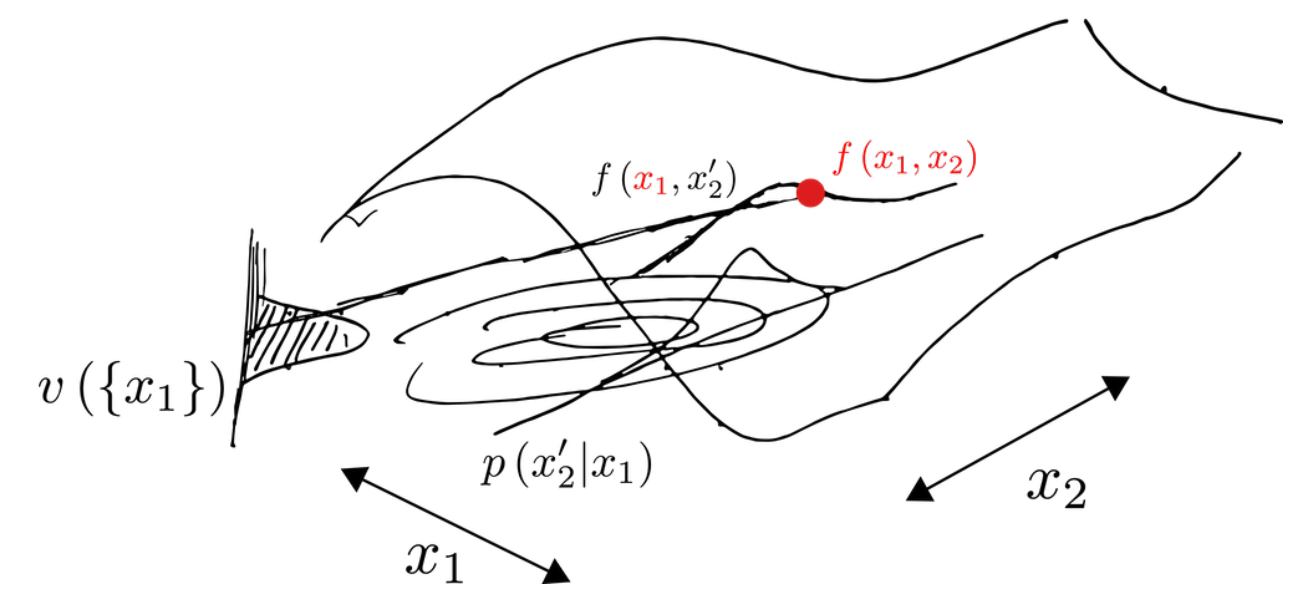
9. A key property is that

$$f(x) = \sum_d \varphi_x(f, d)$$

So the prediction can be recovered from attributions.



10. Here is a geometric interpretation for computing $V_x(\{1\})$



SHAP - Computational

11. While elegant, SHAP seems computationally absurd:

- $v_x(S)$ is a complex conditional expectation
- $S_d(i)$ can include very many subsets

12. There are a few ways to approximate $v_x(S)$

(i) Assume $x_S \perp\!\!\!\perp x_{S^c}$. Then,

$$v_x(S) = \mathbb{E}_{p(x'_{S^c})}\left[f(x_S, x'_{S^c})\right]$$

$$\approx \frac{1}{N}\sum_n f(x_S, x'_{nS^c})$$

sample to explain, at features S

(ii) Learn a new model for $x'_{S^c}|x_S$

(i) Sample

$$x'_{nS^c} \sim p(\cdot|x_S)$$

(ii) Approximate

$$V_x(S) = \sum_{n=1}^{N^*} f(x_S, x'_{nS^c})$$

13. There is connection between Shapley values and a specific type of Weighted Linear Regression. We will work through this in our Demo. The algorithm is called "KernelSHAP."

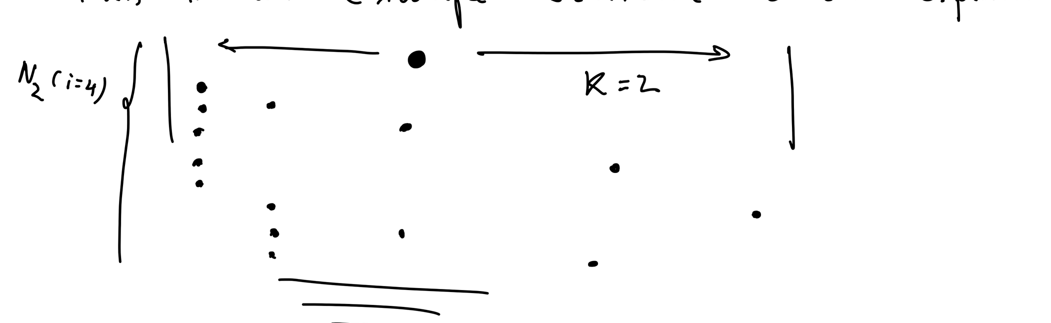
14. Another idea is to prune $S_d(i)$ cleverly. E.g., for a word at the start of a sentence, don't bother with sets of words near the end.

$\Longrightarrow$ We often have distances between features!

15. This is formalized by L and C-Shapley. For example,

$$\varphi_x^L(f, i) = \frac{1}{|N_k(i)|}\sum_{S \in N_k(i)}\frac{1}{\binom{N_k(i) - 1}{|S| - 1}}\left[V_x(S) - V_x(S - \{i\})\right]$$

Where $N_k(i) = \{\text{features w/in distance } K \text{ of feature } i\}$



And the same idea works for images