Model Interpretation

Yifei Sun

Department of Biostatistics
Mailman School of Public Health
Columbia University

Interpreting black-box models

- Global interpretation:
  - **Variable importance**: identify the variables with the largest overall impact
  - **Partial dependence plots**: the typical influence of a feature on the response variable across all observations
  - **Individual conditional expectations**: more information than partial dependence plots
- Local interpretation
  - Given a new observation, what were the most influential variables that determined the predicted outcome?
  - `lime` (Local Interpretable Model-agnostic Explanations)

Partial dependence plots (PDP)

- Graphical renderings of the $f(X)$ as a function of its arguments? Difficult if $p > 3$
- An alternative: partial dependence of $f(X)$ on a selected small subset of the input variables
- Consider the subvector $X_{\mathcal{S}}$ of $l < p$ predictors, indexed by $\mathcal{S} \subset \{1, 2, \ldots, p\}$
- Let $\mathcal{C}$ be the complement set, $\mathcal{C} \cup \mathcal{S} = \{1, 2, \ldots, p\}$
- The partial dependence of $f(X)$ on $X_{\mathcal{S}}$ is

$$f_{\mathcal{S}}(X_{\mathcal{S}}) = E_{X_{\mathcal{C}}} f(X_{\mathcal{S}}, X_{\mathcal{C}})$$

- This can be estimated by

$$\frac{1}{n} \sum_{i=1}^{n} f(X_{\mathcal{S}}, x_{i\mathcal{C}})$$

Partial dependence plots

- Partial dependence functions are not the effect of $X_{\mathcal{S}}$ on $f(X)$ ignoring the effects of $X_{\mathcal{C}}$, i.e.,

$$E(f(X_{\mathcal{S}}, X_{\mathcal{C}}) \mid X_{\mathcal{S}})$$

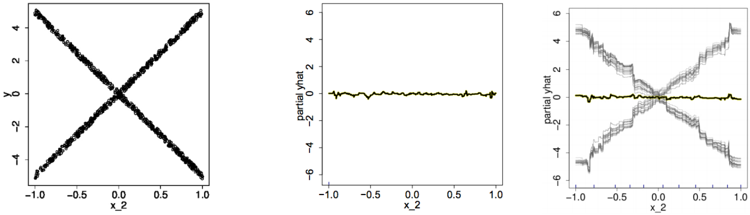
- They are the same only if $X_{\mathcal{S}}$ and $X_{\mathcal{C}}$ are independent
- Example: $f(X) = h_1(X_{\mathcal{S}}) + h_2(X_{\mathcal{C}})$

Partial dependence plots: disadvantages

- The realistic maximum number of features in a partial dependence function is two
- When the features are correlated, we create data points in areas of the feature distribution where the actual probability is very low
- Heterogeneous effects might be hidden

Individual conditional expectations (ICE)

Example: $Y = 0.2X_1 - 5X_2 + 10X_2 \cdot I(X_3 \geq 0) + \epsilon$

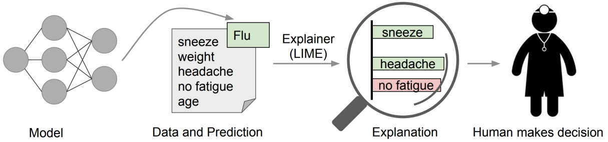


- For each observed $x_{i\mathcal{C}}$, a curve is plotted against the observed values of $X_{\mathcal{S}}$
- Each curve defines the conditional relationship between $X_{\mathcal{S}}$ and f fixed values of $X_{\mathcal{C}}$
- One line represents the predictions for one instance if we vary the feature of interest

Individual conditional expectations

- ICE curves can only display one feature meaningfully
- Some points in the lines might be invalid data points (same problem as PDP)
- Combine individual conditional expectation curves with the partial dependence plot
- One may also consider centered ICE to remove level effects (all curves originate at 0)

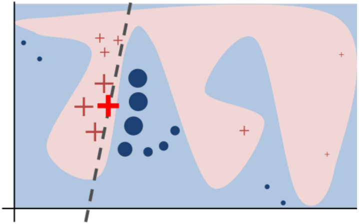
Local interpretation



- Assumption: The complex models are linear on a local scale
- Fit a simple model around a single observation that mimic how the global model behaves at that locality - local surrogate
- The simple model can be used to explain the prediction

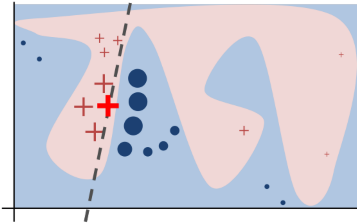
LIME: more details

- The black-box model's decision function f is represented by the background; the bold red cross is the instance being explained
- LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained



LIME: more details

- Define an explanation as a model $g \in G$, where G is a class of potentially interpretable models (e.g., linear models)
- Let $L(f, g, \pi_x)$ be a measure of how unfaithful g is in approximating f in the locality defined by π_x (i.e., a proximity measure)
- The dashed line is the learned explanation that is locally faithful, i.e., $\arg\min_{g \in G}\{L(f, g, \pi_x) + \Omega(g)\}$, where $\Omega(g)$ is a measure of complexity



Functions in `lime`

- `lime()`
  - Creates a list that contains the machine learning model and the feature distributions for the training data
- `explain()`
  - Perform the LIME algorithm
- `plot_features()`
  - Visualization
- Promising method, still in development phase