Explainable AI (CSE615)

Quiz-2, Date: Feb 19, 2025

Max Time: 40 mins Winter 2025 Max marks: 35

1. Consider the model

$$f(x, y) = 2 + 3x + 4y + 6xy,$$

where x and y are independent standard normal random variables, with sample points as following. [5+4]

Observation	x	y
1	-1	-1
2	-1	1
3	0	0
4	1	-1
5	1	1

Table 1: Sample observations for calculating Friedman's H-statistic.

(a) Compute the partial dependence function for x and y. Alternately, show the calculation for 1-2 observations.

(b) Calculate the Friedman H-statistic H^2_{xy}. Use the calculation from partial dependence function here.

Solution - (a) If the PDP of x and y were determined using a subset of the given observations, it will be graded based on the steps, the formula used, and the result. The other method of deriving the PDP for x and y is as follows.

Standard normal random variables have a mean of 0 and a standard deviation of 1.

- **For x:**

$$PD_x(x) = E[f(x, y) \mid x] = 2 + 3x + 4E[y] + 6x\, E[y].$$

Since $E[y] = 0$, we obtain

$$PD_x(x) = 2 + 3x.$$

- **For y:**

$$PD_y(y) = E[f(x, y) \mid y] = 2 + 3E[x] + 4y + 6y\, E[x].$$

Again, since $E[x] = 0$, we have

$$PD_y(y) = 2 + 4y.$$

(b) The overall mean of the model is:

$$E[f(x,y)] = 2 + 3E[x] + 4E[y] + 6E[x]E[y] = 2.$$

Define the centered prediction function as:

$$f^*(x,y) = f(x,y) - E[f(x,y)] = f(x,y) - 2.$$

Thus,

$$f^*(x,y) = 3x + 4y + 6xy.$$

Similarly, the centered partial dependence functions are:

$$\tilde{PD}_x(x) = PD_x(x) - 2 = 3x, \quad \tilde{PD}_y(y) = PD_y(y) - 2 = 4y.$$

The interaction residual is given by subtracting the additive (main effect) parts from f^*:

$$I(x,y) = f^*(x,y) - \Big[\tilde{PD}_x(x) + \tilde{PD}_y(y)\Big].$$

Substituting, we have:

$$I(x,y) = \Big(3x + 4y + 6xy\Big) - \Big(3x + 4y\Big) = 6xy.$$

Since $I(x,y) = 6xy$ and x and y are independent standard normal variables, note that

$$\mathrm{Var}(xy) = E[x^2y^2] - \big(E[xy]\big)^2.$$

Because $E[x^2] = E[y^2] = 1$ and $E[xy] = E[x]E[y] = 0$, we have:

$$E[x^2y^2] = E[x^2]E[y^2] = 1 \quad \Longrightarrow \quad \mathrm{Var}(xy) = 1.$$

Thus,

$$\mathrm{Var}(I(x,y)) = \mathrm{Var}(6xy) = 36 \cdot \mathrm{Var}(xy) = 36.$$

Assuming that the main effects and the interaction term are uncorrelated, we have:

$$\mathrm{Var}\big(f^*(x,y)\big) = \mathrm{Var}(3x + 4y) + \mathrm{Var}(6xy).$$

Since x and y are independent:

$$\mathrm{Var}(3x + 4y) = 9\ \mathrm{Var}(x) + 16\ \mathrm{Var}(y) = 9 + 16 = 25.$$

We already computed $\mathrm{Var}(6xy) = 36$. Therefore,

$$\mathrm{Var}\big(f^*(x,y)\big) = 25 + 36 = 61.$$

$$H_{xy}^2 = \frac{\mathrm{Var}\big(I(x,y)\big)}{\mathrm{Var}\big(f^*(x,y)\big)} = \frac{36}{61}.$$

$$H_{xy} = \sqrt{\frac{36}{61}} = \frac{6}{\sqrt{61}} \approx 0.768.$$

Final Answer: $H_{xy} \approx 0.768$.

b) Alternative solution (without centering)

$$H_{xy}^2 = \frac{\sum_{i=1}^{N}\Big(f(x_i,y_i) - \hat{f}(x_i) - \hat{f}(y_i)\Big)^2}{\sum_{i=1}^{N} f(x_i,y_i)^2}$$

$$\hat{f}(x) = 2 + 3x, \quad \hat{f}(y) = 2 + 4y$$

$$H_{xy}^2 =$$

$$= \frac{(1-(-1)-(-2))^2 + (-3-(-1)-6)^2 + (2-2-2)^2 + (-5-5+2)^2 + (15-5-6)^2}{(1)^2 + (-3)^2 + (2)^2 + (-5)^2 + (15)^2} = \frac{164}{264}$$

Final Answer: $H_{xy} \approx 0.788$.

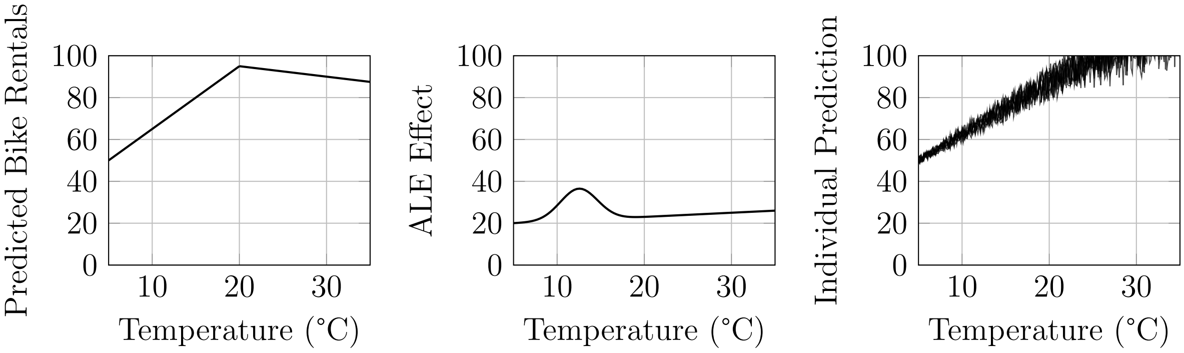
Any combination of above two solution for part (b) are acceptable.

2. Given the Partial Dependence Plot (PDP), Accumulated Local Effects (ALE) plot, and Individual Conditional Expectation (ICE) plot for the feature "Temperature" in a bike rental prediction model, provide an interpretation of these plots. [6]

Solution - 1. Partial Dependence Plot (PDP) -

The PDP illustrates the average effect of "Temperature" on the predicted number of bike rentals across the dataset. The observed trend indicates:

- **Increase from 5°C to 20°C:** A nearly linear rise in predicted rentals suggests that, on average, warmer temperatures within this range encourage more bike rentals.



(a) PDP

(b) ALE Plot

(c) ICE Plot

Figure 1: Model interpretation plots for the feature “Temperature”.

- **Plateau and Slight Decline beyond 20°C:** The flattening and slight decrease imply that extremely warm temperatures may not further boost rentals and could even deter biking activity, possibly due to discomfort in excessive heat.

This pattern reflects the general preference of users for moderate temperatures when choosing to rent bikes.

2. Accumulated Local Effects (ALE) Plot -

The ALE plot provides a localized understanding of ”Temperature’s” impact by accounting for the feature’s distribution and interactions. Key observations include:

- **Monotonic Increase Across All Temperatures:** Unlike the PDP, the ALE plot shows a continuous rise, indicating that, conditionally, increasing temperatures consistently elevate the prediction of bike rentals.
- **Bump Between 9°C and 18°C:** This suggests that within this specific range, temperature has a more pronounced effect on rental counts, possibly due to optimal biking conditions. Where in first half there is steep increase, while in second half it decreases to match the similar monotonic trend as the whole graph.

The differences between the ALE and PDP plots may arise from the ALE plot’s consideration of the actual data distribution, providing a nuanced view that mitigates the influence of regions with sparse data.

3. Individual Conditional Expectation (ICE) Plot -

The ICE plot displays the relationship between ”Temperature” and predicted rentals for individual instances, revealing:

- **Variability Among Instances:** Some lines show a sharp increase with temperature, while others remain flat or decline at higher temperatures, indicating that the effect of temperature varies across different conditions or user profiles.
- **Potential Interactions:** The heterogeneity suggests interactions between "Temperature" and other features (e.g., humidity, time of day) influencing rental behavior.

This variability highlights that while temperature is a significant factor, its impact is modulated by other contextual factors.

Reconciling the Plots -

The PDP provides a global average effect, potentially smoothing out individual variations and interactions. The ALE plot offers a more localized perspective, accounting for the feature's distribution and reducing bias from correlated features. The ICE plot uncovers individual differences, emphasizing the importance of considering interactions and conditional effects.

The slight decline in the PDP beyond 20°C, contrasted with the continuous increase in the ALE plot, suggests that while the overall trend is positive, certain subsets of data (perhaps underrepresented in the ALE due to its local nature) experience a decrease in rentals at higher temperatures. This discrepancy underscores the importance of using multiple interpretative tools to gain a comprehensive understanding of feature effects.

In summary, "Temperature" positively influences bike rental predictions, but its effect is complex and context-dependent. Analyzing PDP, ALE, and ICE plots collectively allows for a deeper insight into both the general trends and individual variations, facilitating more informed decision-making and model refinement.

3. Consider a dataset with a categorical feature "Color" having three possible values: Red, Blue, Green. LIME perturbs this feature by randomly sampling new values from this set. How might this perturbation strategy lead to biases in the explanation? Suggest a better perturbation strategy for categorical features. [6]

A. 1. Loss of Contextual Relationships: In many datasets, categorical features aren't independent—they may be correlated with other features. Example: If Red cars are more common in luxury models, replacing "Red" with "Blue" randomly may create unrealistic samples. This leads to LIME learning an incorrect local decision boundary.

2. Unequal Class Distribution Issues. If the original dataset has 90% Red, 5% Blue, 5% Green, but LIME perturbs the feature uniformly, it introduces a bias. The model

might rarely see Blue in real data, but LIME generates too many Blue instances, making the explanation unreliable.

To generate more realistic perturbed instances, we can use frequency-aware sampling. For correlated features, ensure perturbations do not introduce an unrealistic pairing.

4. You apply LIME to explain a deep learning model for credit scoring and get different explanations for the same instance when you rerun LIME multiple times. What might be causing this instability? How would you modify LIME to ensure more consistent explanations? [6]

   A.

   - Random Perturbation of Data. If the perturbation process is too random or not controlled properly, different samples can lead to different surrogate models, changing the feature importance scores.
   - Choice of Kernel Width across different runs.
   - If the number of perturbed samples is too low, the surrogate model is fitted on an insufficient dataset, leading to variations across runs
   - If features are highly correlated, small perturbations can cause large changes in predictions, leading to different explanations each time LIME is run

   To improve stability,

   - Increase the Number of Perturbed Samples
   - Optimize kernel width
   - Run LIME Multiple Times and Aggregate Results

5. Consider a binary classification problem where we use LIME to explain a black-box model's prediction for a given instance. Let $X_0 = (5.0, 2.5)$ be a data point. We generate the following perturbed samples by adding Gaussian noise.

Perturbed Instance	(X_1, X_2)
X_1	(4.8, 2.3)
X_2	(5.2, 2.6)
X_3	(4.5, 2.0)
X_4	(5.5, 2.8)
X_5	(6.0, 3.2)

The σ value for the kernel function is set to 0.75.
Using $w(X_i) = \exp\left(-\frac{||X_i - X_0||^2}{\sigma^2}\right)$ where $||X_i - X_0||$ represents the Euclidean distance between X_i and X_0. [8]

(a) Compute the Euclidean distances between the perturbed instances and X_0.

(b) Calculate the weights for each perturbed instance using the given kernel function.

(c) Determine which instances you would choose for training a local model and justify your selection.

Sol: Using the Euclidean distance formula:

$$||X_i - X_0|| = \sqrt{(X_1 - 5.0)^2 + (X_2 - 2.5)^2}$$

Instance X_i	Distance d_i
(4.8, 2.3)	0.283
(5.2, 2.6)	0.224
(4.5, 2.0)	0.707
(5.5, 2.8)	0.583
(6.0, 3.2)	1.22

$$w(X_i) = \exp\left(-\frac{d_i^2}{0.75^2}\right) = \exp\left(-\frac{d_i^2}{0.5625}\right)$$

Instance X_i	Distance d_i	Weight $w(X_i)$
(4.8, 2.3)	0.283	0.867
(5.2, 2.6)	0.224	0.914
(4.5, 2.0)	0.707	0.411
(5.5, 2.8)	0.583	0.546
(6.0, 3.2)	1.22	0.070

We select instances with higher weights (above 0.7):

- **Chosen Instances:** (4.8, 2.3), (5.2, 2.6)
- **Less Preferable:** (4.5, 2.0), (5.5, 2.8), (6.0, 3.2) (too far)

Justification:

- The chosen instances have the highest weights, meaning they are closest to X_0 in feature space and contribute more to the local model.
- The instance (6.0, 3.2) has a very low weight (0.070), making it much less relevant for explaining X_0.
- Selecting closer points ensures that the local surrogate model better represents the decision boundary around X_0.