Explainable AI (CSE615)

Midsem Exam, Date: Feb 28, 2025

Max Time: 1.5 hrs | Winter 2025 | Max marks: 80

1. A machine learning model predicts house prices (y) based on two features - square footage (x_1) and number of bedrooms (x_2). The model's prediction function is as follows.

   $f(x_1, x_2) = 50 + 2x_1 + 10x_2$

   A homeowner has a 1200 sq. ft. house with 3 bedrooms, and the predicted price is as follows.

   $f(1200, 3) = 50 + 2 * 1200 + 10 * 3 = 2480$

   However, the homeowner wants to see how they could increase the predicted price to 2700. Use Wachter's counterfactual generation method to find the counterfactual values p, q, such that $f(p, q) = 2700$. Use the Euclidean distance (you can use d^2 instead of d to avoid square root calculations) and solve the arg min optimization. As part of this process (to minimize), you would have to take the partial derivative of p and the partial derivative of q and set them to zero. Assume λ to be 0.01. Comment on the obtained counterfactual values p and q and whether or not they are helpful in increasing the house price. [20]

   Answer

   Prediction function: $f(x_1, x_2) = 50 + 2x_1 + 10x_2$

   Original instance: f(1200, 3) = 2480

   Target prediction: $f(p, q) = 2700$

   $\lambda = 0.01$

   Loss function to minimize (Wachter's formulation with squared Euclidean distance), $L(p, q) = (p - 1200)^2 + (q - 3)^2 + \lambda \cdot (f(p, q) - 2700)^2$

   Plugging in $f(p, q) = 50 + 2p + 10q$, we get,

   $L(p, q) = (p - 1200)^2 + (q - 3)^2 + \lambda \cdot (50 + 2p + 10q - 2700)^2 = (p - 1200)^2 + (q - 3)^2 + \lambda \cdot (2p + 10q - 2650)^2$

   Compute Partial Derivatives. We take partial derivatives with respect to p and q and set them to 0.

   Partial derivative w.r.t p: $\frac{\partial L}{\partial p} = 2(p\text{-}1200) + \lambda \cdot 2(2p + 10q - 2650) \cdot 2 = 2(p\text{-}1200) + 4\lambda \cdot (2p + 10q - 2650)$

   Partial derivative w.r.t q: $\frac{\partial L}{\partial q} = 2(q\text{-}3) + \lambda \cdot 2(2p + 10q - 2650) \cdot 10 = 2(q\text{-}3) + 20\lambda \cdot (2p + 10q - 2650)$

   Set the derivatives to zero. Set both partial derivatives to zero and substitute $\lambda = 0.01$.

   i) $2(\text{p}-1200) + 0.04 \cdot (2p + 10q - 2650) = 0$

ii) 2(q−3) $+ 0.2 \cdot (2p + 10q - 2650) = 0$

From i), 2p - 2400 + 0.08p + 0.4q - 106 = 0, i.e., 2.08p + 0.4q = 2506

From ii), 2q - 6 + 0.4p + 2q - 530 = 0, i.e., 0.4p + 4q = 536

p = (536 - 4q)/0.4 = 1340 - 10q. Substitute this value into the other equation.

2.08 (1340 - 10q) + 0.4q = 2506, i.e., 2787.2 - 20.8q + 0.4q = 2506, i.e., 281.2 = 20.4q, i.e., q = 13.78

p = 1340 - 10q = 1340 - 137.8 = 1202.2

q = 13.78

2. You are given a dataset where LIME is used to explain a prediction at $X = (X_1, X_2)$ = (40, 180) for a regression model. The perturbed data points are as follows. [15]

Instance	**Age(X_1)**	**Cholesterol(X_2)**	**Model Prediction F(x)**
A	38	175	0.72
B	42	185	0.80
C	50	190	0.87
D	35	170	0.65
E	45	200	0.82

Using the exponential kernel function,

- Compute the weights for $\sigma = 5$ and $\sigma = 15$
- Compare how the choice of σ affects the importance of perturbed points in the explanation
- Would a very large σ be beneficial for explaining this model? Why or why not?

Answer

Weighting function, $w(x) = e^{\frac{-D(x,x_o)^2}{\sigma^2}}$, where x_o is the original instance to be explained and x is the perturbed instance.

Here, $x_o = X_o$ = (40, 180)

For A, $D(X_A, X_o)^2 = (38 - 40)^2 + (175 - 180)^2 = 4 + 25 = 29$

For B, $D(X_B, X_o)^2 = (42 - 40)^2 + (185 - 180)^2 = 4 + 25 = 29$

For C, $D(X_C, X_o)^2 = (50 - 40)^2 + (190 - 180)^2 = 100 + 100 = 200$

For D, $D(X_D, X_o)^2 = (35 - 40)^2 + (170 - 180)^2 = 25 + 100 = 125$

For E, $D(X_E, X_o)^2 = (45 - 40)^2 + (200 - 180)^2 = 25 + 400 = 425$

Kernel weights for $\sigma = 5$

For A, $w_A(x) = e^{-29/25} = e^{-1.16} = 0.313$

For B, $w_B(x) = e^{-29/25} = e^{-1.16} = 0.313$

For C, $w_C(x) = e^{-200/25} = e^{-8} = 0.00033$

For D, $w_D(x) = e^{-125/25} = e^{-5} = 0.0067$

For E, $w_E(x) = e^{-425/25} = e^{-17} = 0.00000004$

Kernel weights for $\sigma = 15$

For A, $w_A(x) = e^{-29/225} = e^{-0.1288} = 0.8790$

For B, $w_B(x) = e^{-29/225} = e^{-0.1288} = 0.8790$

For C, $w_C(x) = e^{-200/225} = e^{-0.8888} = 0.4111$

For D, $w_D(x) = e^{-125/225} = e^{-0.5555} = 0.5737$

For E, $w_E(x) = e^{-425/225} = e^{-1.8888} = 0.1512$

Comparison

- For $\sigma = 5$, instances C, D, and E have very small weights (close to 0). Only A and B have meaningful contributions. Narrow kernel width gives high importance to very close points but ignores distant ones.
- For $\sigma = 15$, all instances now have significant weights. Even E, which is far away, contributes 15%. Large σ includes more distant points in the explanation, leading to a more global effect.

Small σ is better for highly nonlinear models. It focuses only on the most relevant local points. Distant points are ignored. Large σ is better for smoother models. It takes a broader view of how the model behaves.

3. A function $F(x)$ is said to have an interaction between three (numeric) variables x_j, x_k and x_l, if

$$E_x\left[\frac{\partial^3 F(x)}{\partial x_j \partial x_k \partial x_l}\right]^2 > 0 \tag{1}$$

If there is no such three-variable interaction, $F(x)$ can be expressed as a sum of three functions, each independent of one of the three variables

$$F(x) = f_{\backslash j}(x_{\backslash j}) + f_{\backslash k}(x_{\backslash k}) + f_{\backslash l}(x_{\backslash l}) \tag{2}$$

Here $x_{\backslash j}$, $x_{\backslash k}$ and $x_{\backslash l}$ each respectively represent all of the variables except x_j, x_k and x_l. Given this description, deduce Friedman's H-statistic for three variables.

Also, find the relations (simplifying Friedman's H-statistic), (i) when variables x_j, x_k and x_l do not impact the result through simultaneous interaction with each other, and (ii) when x_j, x_k and x_l only effect through mutual interactions. [15]

Answer

We know that, if a given variable x_j interacts with none of the other variables, then the function can be expressed as

$$F(x) = f_j(x_j) + f_{\backslash j}(x_{\backslash j}) \tag{3}$$

And if there is no three variable interaction, it can be measured as

$$F(x) = f_{\backslash j}(x_{\backslash j}) + f_{\backslash k}(x_{\backslash k}) + f_{\backslash l}(x_{\backslash l}) \tag{4}$$

If variables x_j, x_k and x_l do not participate in a joint three-variable interaction, then from Equation 4, the partial dependence of $F(x)$ on these three variables can be expressed in terms of the respective lower order partial dependencies as

$$F_{jkl}(x_{ij}, x_{ik}, x_{il}) = F_{jk}(x_{ij}, x_{ik}) + F_{jl}(x_{ij}, x_{il}) + F_{kl}(x_{ik}, x_{il}) - F_j(x_{ij}) - F_k(x_{ik}) - F_l(x_{il}) \tag{5}$$

We know that the h statistic is normalized variance of difference between, when n-variable interact simulataneously, and when they do not. i.e.

$$H^2_{jkl} = \frac{\sum_{i=1}^{N} [\hat{F}_{jkl}(x_{ij}, x_{ik}, x_{il}) - \hat{F}_{jk}(x_{ij}, x_{ik}) - \hat{F}_{jl}(x_{ij}, x_{il}) - \hat{F}_{kl}(x_{ik}, x_{il}) + \hat{F}_j(x_{ij}) + \hat{F}_k(x_{ik}) + \hat{F}_l(x_{il})]^2}{\sum_{i=1}^{N} \hat{F}^2_{jkl}(x_{ij}, x_{ik}, x_{il})} \tag{6}$$

where,

$$\hat{F}_s(x_s) = \frac{1}{N} \sum_{i=1}^{N} F(x_s, x_{i \setminus s}) \tag{7}$$

(i) From Equation 5, we can say that

$$\begin{aligned} H^2_{jkl} = & \sum_{i=1}^{N} [\hat{F}_{jk}(x_{ij}, x_{ik}) + \hat{F}_{jl}(x_{ij}, x_{il}) + \hat{F}_{kl}(x_{ik}, x_{il}) - \hat{F}_j(x_{ij}) - \hat{F}_k(x_{ik}) - \hat{F}_l(x_{il}) \\ & - \hat{F}_{jk}(x_{ij}, x_{ik}) - \hat{F}_{jl}(x_{ij}, x_{il}) - \hat{F}_{kl}(x_{ik}, x_{il}) + \hat{F}_j(x_{ij}) + \hat{F}_k(x_{ik}) + \hat{F}_l(x_{il})]^2 / (\sum_{i=1}^{N} \hat{F}^2_{jkl}(x_{ij}, x_{ik}, x_{il})) = 0 \end{aligned}$$

(ii)

$$H^2_{jkl} = \frac{\sum_{i=1}^{N} [F_{jkl}(x_{ij}, x_{ik}, x_{il})]^2}{\sum_{i=1}^{N} F^2_{jkl}(x_{ij}, x_{ik}, x_{il})} = 1 \tag{8}$$

4. For the following scenarios, explain which among PDP, ALE or ICE plot would you use for interpreting the results of the model. More than one plot can also be used. [15]

   a) **Scenario-1** is about understanding the effect of marketing spend on sales. A company builds a sales prediction model using the following features.

      i) Marketing spend: Continuous variable representing the advertising budget in dollars

      ii) Store location: Categorical variable indicating different regions (e.g., Urban, Suburban, Rural)

      Assume that these two features are weakly correlated. The business team wants to understand

      - How marketing spend affects sales on average (to allocate budget effectively)
      - Whether spending more on advertisements consistently increases sales or if there is a saturation point

   b) **Scenario-2** A model predicting student exam performance based on

      i) Study time

      ii) IQ score

      We would like to study the impact of these variables on the students.

c) **Scenario-3** A model predicting employee attrition uses

   i) Job Level (categorical; [Entry, Mid, Senior])

   ii) Years at company (continuous variable)

   These two features are correlated.

Answer

For Scenario-1,

- Since the features are weakly correlated, PDP can be used to understand average effects and saturation trends.
- ICE can be used to detect heterogeneity in effects across data instances.

For Scenario-2, If study time and IQ score are correlated,

- ALE can be used to understand the general, unbiased effects of study time and IQ.
- ICE plots can be used to show how study time or IQ affects specific students. Use ICE to see how individual students are impacted by changes in those features.

If study time and IQ are correlated, PDP plots can be used for average trends.

For Scenario-3,

- Since the two features are correlated, ALE plots can be used. It provides unbiased local effects. For Years at company, ALE avoids extrapolating into unrealistic combinations (e.g., Entry-level employees with 20 years).
- ICE plots are useful to uncover heterogeneity across employees (e.g., whether long-tenured Entry-level employees are more likely to leave).

5. A medical dataset contains the following attributes for 10 patients. [15]

   a) Temperature (Low, Normal, High)

   b) Cough (No, Mild, Severe)

   c) Fatigue (No, Yes)

   d) Flu Diagnosis (Yes, No)

Using the OneR algorithm, determine the best single-feature decision rule for predicting whether a patient has flu. Show all calculations, including the computation of the error rates. Write the decision rule involving the best selected feature.

Answer

Step 1: Evaluate **Temperature**

- Low: Patients 3, 6, 9 → Flu: No, Yes, No. Majority = **No**. Errors = 1 (Patient 6).
- Normal: Patients 1, 5, 8 → Flu: No, No, No. Majority = **No**. Errors = 0.

Patient	Temperature	Cough	Fatigue	Flu Diagnosis
1	Normal	Mild	No	No
2	High	Severe	Yes	Yes
3	Low	No	No	No
4	High	Severe	No	Yes
5	Normal	Mild	No	No
6	Low	Mild	Yes	Yes
7	High	Severe	Yes	Yes
8	Normal	Mild	No	No
9	Low	No	Yes	No
10	High	Mild	Yes	No

- High: Patients 2, 4, 7, 10 → Flu: Yes, Yes, Yes, No. Majority = **Yes**. Errors = 1 (Patient 10).

Total Errors for Temperature: 1 + 0 + 1 = 2

Step 2: Evaluate **Cough**

- No: Patients 3, 9 → Flu: No, No. Majority = **No**. Errors = 0.
- Mild: Patients 1, 5, 8, 10 → Flu: No, No, No, No. Majority = **No**. Errors = 0.
- Severe: Patients 2, 4, 7 → Flu: Yes, Yes, Yes. Majority = **Yes**. Errors = 0.

Total Errors for Cough: 0 + 0 + 0 = 0

Step 3: Evaluate **Fatigue**

- No: Patients 1, 3, 4, 5 → Flu: No, No, Yes, No. Majority = **No**. Errors = 1 (Patient 4).
- Yes: Patients 2, 6, 7, 9, 10 → Flu: Yes, Yes, Yes, No, No. Majority = **Yes**. Errors = 2 (Patients 9 and 10).

Total Errors for Fatigue: 1 + 2 = 3

Conclusion

- Temperature: 2 errors
- Cough: **0 errors** (Best attribute)
- Fatigue: 3 errors

Final OneR Rule Based on Cough

- If Cough = No, then Flu = No
- If Cough = Mild, then Flu = No
- If Cough = Severe, then Flu = Yes

Error Rate: 0/10 = 0%