Explainable AI (CSE615)

End-Sem, Date: May 4, 2025

Max Time: 2hrs | Winter 2025 | Max marks: 70

1. McGrath [2018] proposed updated distance metric for the Wachter method (Equation 1), where they added additional parameter θ_j to assign different weights to different features. The values of θ_j is calculated via global feature importance or Nearest Neighbors approach.

$$d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j} \theta_j \tag{1}$$

   (a) What are the issues this new method solves as compared to Wachter method.

   (b) Does Dandl method solves similar issues? If yes, what objective in the overall Dandl method resolve this.

   (c) How are both method different, (using same objective mentioned in (b)), i.e., examples where one of Dandl and McGrath solves the problem while other does not.

[3+2+4]

Answer

   (a) The method associate different weights to different features, which results in better data point generation, as some features like salary or loan amount can have high variations while features like age cannot have much variation. Also features like gender for a person A cannot change which can easily be done in this case by setting $\theta = 0$ for that feature.

   (b) Yes, In Dandl method, objective 4,

$$o_4(x, X_{\text{obs}}) \;=\; \frac{1}{p} \sum_{j=1}^{p} \delta_G\big(x'_j,\, x_j^{[1]}\big) \tag{2}$$

   where $x^{[1]} \in X_{obs}$, is working on similar principle, where the objective is that the new data point is in the similar distribution as the training data.

   (c) The Dandl method could have the data point in similar space but the features like gender, which are rare to change for a datapoint, may also exist. So McGrath might work better for this objective.

2. Consider a simple Convolutional Neural Network where the final convolutional layer produces the following two feature maps (after applying ReLU), each of size 2x2.

   Feature Map 1: $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$

   Feature Map 2: $\begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix}$

These feature maps are followed by global average pooling, and the output is passed to a fully connected layer for classification into one class. Suppose the weights connecting the pooled values to this class are as follows.

- Weight for feature map 1: $w_1 = 0.5$
- Weight for feature map 2: $w_2 = 1.0$

Given this, compute the following.

a) The global average pooled value for each feature map.

b) The class score before softmax.

c) The Class Activation Map (CAM) for this class by combining the feature maps with their corresponding weights.

[4+2+4]

Answer

a) Global average pooling (mean of 2×2 entries)
Map 1: $(1 + 2 + 0 + 1)/4 = 1.0$
Map 2: $(0 + 1 + 2 + 1)/4 = 1.0$

b) Class score (before softmax)
Score=$0.5\times1.0+1.0\times1.0=1.5$

c) CAM = (0.5 × Feature map 1) + (1.0 × Feature map 2)
$= 0.5 * \begin{bmatrix}1 & 2\\0 & 1\end{bmatrix} + 1.0 * \begin{bmatrix}0 & 1\\2 & 1\end{bmatrix} = \begin{bmatrix}0.5 & 1.0\\0.0 & 0.5\end{bmatrix} + \begin{bmatrix}0.0 & 1.0\\2.0 & 1.0\end{bmatrix} = \begin{bmatrix}0.5 & 2.0\\2.0 & 1.5\end{bmatrix}$

3. Consider a CNN that produces the following 3 feature maps (A_1, A_2, A_3) from its last convolutional layer. Each is of size 2×2.

$$A_1 = \begin{bmatrix}1 & 2\\0 & 1\end{bmatrix}$$

$$A_2 = \begin{bmatrix}0 & 1\\2 & 1\end{bmatrix}$$

$$A_3 = \begin{bmatrix}1 & 0\\1 & 1\end{bmatrix}$$

For a particular class (say "dog"), the gradient of the class score, y^c, with respect to each feature map has been computed at the corresponding activation points. These gradients are as follows.

$$\frac{\partial y^c}{\partial A_1} = \begin{bmatrix}0.1 & 0.3\\0.2 & 0.4\end{bmatrix}$$

$$\frac{\partial y^c}{\partial A_2} = \begin{bmatrix}0.2 & 0.1\\0.3 & 0.2\end{bmatrix}$$

$$\frac{\partial y^c}{\partial A_3} = \begin{bmatrix}0.4 & 0.5\\0.1 & 0.0\end{bmatrix}$$

Given this, compute the following.

a) The average gradient (importance weight) for each feature map.

b) The Grad-CAM heatmap as a weighted sum of the feature maps using the importance weights.

c) Apply ReLU to the Grad-CAM heatmap.

d) Briefly discuss the possible interpretation of the heatmap.

[3+4+1+2]

Answer

a) Use global average pooling over gradients.
$\alpha_1 = (0.1+0.3+0.2+0.4)/4 = 0.25$
$\alpha_2 = (0.2+0.1+0.3+0.2)/4 = 0.2$
$\alpha_3 = (0.4+0.5+0.1+0.0)/4 = 0.25$

b) Grad-CAM heatmap $= \alpha_1 * A_1 + \alpha_2 * A_2 + \alpha_3 * A_3$
$= 0.25 * \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} + 0.2 * \begin{bmatrix} 0 & 1 \\ 2 & 1 \end{bmatrix} + 0.25 * \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$
$= \begin{bmatrix} 0.25 & 0.5 \\ 0.0 & 0.25 \end{bmatrix} + \begin{bmatrix} 0.0 & 0.2 \\ 0.4 & 0.2 \end{bmatrix} + \begin{bmatrix} 0.25 & 0.0 \\ 0.25 & 0.25 \end{bmatrix}$
$= \begin{bmatrix} 0.5 & 0.7 \\ 0.65 & 0.7 \end{bmatrix}$

c) Apply ReLU. Since all the values are greater than 0 (no negative values), the heatmap remains unchanged, i.e.,
Grad-CAM heatmap $= \begin{bmatrix} 0.5 & 0.7 \\ 0.65 & 0.7 \end{bmatrix}$

d) This heatmap shows which spatial locations in the final convolutional layer contributed most to the prediction of the class "dog". The highest activation is at (1,2) and (2,2), both at 0.7, indicating that these spatial regions were most influential. These regions would be upsampled and overlaid on the input image to visually highlight where the network was *looking*.

4. Consider a black-box model trained for loan approval. A bank compliance team wants to understand how the model works for fairness audits.

(a) Explain with an example how a **local explanation** using LIME might mislead when interpreted globally.

(b) Describe two challenges of generalizing **local surrogate models** like LIME to obtain global explanations.

(c) Suppose you are given SHAP explanations for 500 loan applicants. How might you aggregate them to extract a global understanding of model behavior? Briefly outline your approach and its limitations.

[3+3+4]

Answer

(a) A local explanation using LIME provides insights into how the model behaves for a single instance. For instance, if a loan applicant is rejected due to low income, LIME might highlight income as the most important factor for that rejection. However, if we generalize this local explanation to the entire dataset, it may be misleading, as income might not be the most important feature in the broader model, which could consider a complex interaction of other features, like credit history or loan amount.

(b) (a) **Instability across different instances:** Local models like LIME may generate different explanations for different instances, making it difficult to derive a consistent global explanation.

(b) **Limited feature interactions:** Local surrogate models fail to capture feature interactions that may emerge when considering the full dataset, leading to a less accurate global explanation.

(c) To aggregate SHAP values for 500 loan applicants, we can:

(a) Compute the **mean absolute SHAP value** for each feature across all applicants to identify globally important features.

(b) Use visualizations like **boxplots** or **bar charts** to display the distribution of SHAP values for each feature.

(c) **Cluster applicants** based on similar SHAP value patterns to capture subgroups with distinct behaviors.

However, this approach assumes that SHAP values provide a global representation, which may not fully capture feature interactions or subgroup-specific behaviors, thus limiting the explanation's completeness.

5. You are working with a CNN trained to classify chest X-ray images into Healthy, Pneumonia, and COVID-19. To make its decisions explainable, use mapping networks and an ontology. For a chest X-ray image, img123, the CNN predicts that it is a COVIDLung and outputs the latent activation vector, $f = [0.4, 0.5, 0.7]$. You are also given concept vectors for the following clinical features.

- GroundGlassOpacities: [0.3, 0.6, 0.7]
- NoPleuralEffusion: [0.4, 0.5, 0.7]
- Consolidation: [0.6, 0.2, 0.4]
- NormalBronchialPattern: [0.1, 0.9, 0.3]
- BilateralDistribution: [0.5, 0.4, 0.6]

The mapping score for each concept, c_i, is computed using cosine similarity.

$$\text{Score}(f, c_i) = \frac{f.c_i}{||f||.||c_i||}$$

$f.c_i$ is the dot product, calculated by multiplying the corresponding components of the two vectors and summing the results. $||f||$ is the magnitude (Euclidean norm) of vector f, found by taking the square root of the sum of the squares of its components.

[**Part A**] Represent the following using Description Logics.

i) A HealthyLung is a normal bronchial pattern with no ground glass opacities or consolidation.

ii) A COVIDLung is a ground glass opacities with bilateral distribution and no pleural effusion.

iii) A PneumoniaLung is a consolidation without bilateral distribution.

[**Part B**]

i) Compute the cosine similarity between f and each concept vector. Report the score for all 5 concepts.

ii) Identify the concepts whose similarity exceeds the threshold of 0.95. These will be treated as relevant concepts.

iii) Write the ABox statements for img123 based on these relevant concepts.

iv) Use the TBox and ABox statements to construct an explanation for the most likely diagnosis (COVIDLung).

[6+20]

Answer.
[**Part A**]

i) HealthyLung $\sqsubseteq$ NormalBronchialPattern $\sqcap$ $\neg$GroundGlassOpacities $\sqcap$ $\neg$Consolidation

ii) COVIDLung $\sqsubseteq$ GroundGlassOpacities $\sqcap$ BilateralDistribution $\sqcap$ NoPleuralEffusion

iii) PneumoniaLung $\sqsubseteq$ Consolidation $\sqcap$ $\neg$BilateralDistribution

[**Part B**]

i) Cosine Similarity
$||f|| = \sqrt{0.4^2 + 0.5^2 + 0.7^2} = \sqrt{0.16 + 0.25 + 0.49} = \sqrt{0.9} = 0.9486$
GroundGlassOpacities, c_1**:** [0.3, 0.6, 0.7]
$f.c_1$ = 0.4*0.3 + 0.5*0.6 + 0.7*0.7 = 0.12 + 0.30 + 0.49 = 0.91
$||c_1|| = \sqrt{0.3^2 + 0.6^2 + 0.7^2} = \sqrt{0.09 + 0.36 + 0.49} = \sqrt{0.94} = 0.9695$
Score = 0.91/(0.9486*0.9695) = 0.91/0.9196 = 0.9895
NoPleuralEffusion, c_2**:** [0.4, 0.5, 0.7]
$f.c_2$ = 0.4*0.4 + 0.5*0.5 + 0.7*0.7 = 0.16 + 0.25 + 0.49 = 0.9
$||c_2|| = \sqrt{0.4^2 + 0.5^2 + 0.7^2} = \sqrt{0.9} = 0.9486$
Score = 0.9/(0.9486*0.9486) = 0.9/0.9 = 1
Consolidation, c_3**:** [0.6, 0.2, 0.4]
$f.c_3$ = 0.4*0.6 + 0.5*0.2 + 0.7*0.4 = 0.24 + 0.1 + 0.28 = 0.62
$||c_3|| = \sqrt{0.6^2 + 0.2^2 + 0.4^2} = \sqrt{0.36 + 0.04 + 0.16} = \sqrt{0.56} = 0.7483$
Score = 0.62/(0.9486*0.7483) = 0.8734
NormalBronchialPattern, c_4**:** [0.1, 0.9, 0.3]
$f.c_4$ = 0.4*0.1 + 0.5*0.9 + 0.7*0.3 = 0.04 + 0.45 + 0.21 = 0.7
$||c_4|| = \sqrt{0.1^2 + 0.9^2 + 0.3^2} = \sqrt{0.01 + 0.81 + 0.09} = \sqrt{0.91} = 0.9539$
Score = 0.7/(0.9486*0.9539) = 0.7735
BilateralDistribution, c_5**:** [0.5, 0.4, 0.6]
$f.c_5$ = 0.4*0.5 + 0.5*0.4 + 0.7*0.6 = 0.20 + 0.20 + 0.42 = 0.82
$||c_5|| = \sqrt{0.5^2 + 0.4^2 + 0.6^2} = \sqrt{0.25 + 0.16 + 0.36} = \sqrt{0.77} = 0.8774$
Score = 0.82/(0.9486*0.8774) = 0.9852

ii) GroundGlassOpacities, NoPleuralEffusion, and BilateralDistribution have a cosine similarity above the thresold of 0.95.

iii) GroundGlassOpacities(img123), NoPleuralEffusion(img123), and BilateralDistribution(img123) are the ABox statements.

iv) In the TBox, we have COVIDLung $\sqsubseteq$ GroundGlassOpacities $\sqcap$ BilateralDistribution $\sqcap$ NoPleuralEffusion. From the ABox, we have GroundGlassOpacities(img123) $\sqcap$ NoPleuralEffusion(img123) $\sqcap$ BilateralDistribution(img123). From these two, we can infer COVIDLung(123).

Explanation/Justification: The image indicates COVID-19 since there are bilateral ground glass opacities and no pleural effusion.