Chapter23 Explainable Machine Learning

1. Explainable ML
   a) Local Explanation : Why do you think this image is a cat
   b) Global Explanation : What do you think a cat looks like
2. Why we need Explainable ML
   a) Curriculum Vitae Filtering
   b) Financial Decision
   c) Model Diagnosis
3. My Point of View
   a) Goal of ML Explanation ≠ you completely know how the ML model work
   b) Human brain is also a black box, but you believe in human
   c) Goal of ML Explanation is make people (customer, boss, yourself) comfortable
4. Interpretable v.s. Powerful
   a) Some models are intrinsically interpretable, but not very powerful
   b) Deep network is difficult to interpret, but it is more powerful
5. Local Explanation : Explain the Decision
   a) Basic Idea
      i) Removing or modifying the value of the components, observing the change of decision
      ii) Large decision change -> important component
   b) Saliency Map
      $\{x_1, \dots, x_n, \dots, x_N\} \rightarrow \{x_1, \dots, x_n + \Delta x, \dots, x_N\}$
      $y_k \rightarrow y_k + \Delta y$
      Compute $\left|\frac{\Delta y}{\Delta x}\right| \rightarrow |\frac{\partial y_k}{\partial x_n}|$
   c) Limitation of Gradient-based Approaches : Gradient Saturation
      To deal with this problem : Integrated gradient, DeepLIFT
   d) Attack Interpretation
      The noise is small, and do not change the classification result
6. Global Explanation : Explain the whole model
   a) Activation Maximization
      i) Find the image that maximizes class probability and also looks like a digit
      ii) $x^* = \arg max\, y_i + R(x)$
          $R(x) = \sum_{i,j} |x_{ij}|$
      iii) With several regularization terms and hyperparameter tuning
   b) Constraint from Generator
      $x = G(z)$ $x^* = \arg max\, y_i \rightarrow z^* = \arg max\, y_i$
      Show image : $x^* = G(z^*)$
7. Using a model to explain another
   a) Some models are easier to interpret
      Using interpretable model to mimic uninterpretable models
   b) Linear model cannot mimic neural network
      However, it can mimic a local region

c) LIME Local Interpretable Model-Agnostic Explanations

i) Given a data point you want to explain

ii) Sample at the nearby

iii) Fit with linear model (or other interpretable models)

iv) Interpret the linear model

v) Application on Image

Sample at the nearby

Randomly delete some segments

Fit with linear model

$x_1, \dots, x_m, \dots, x_M$ M is the number of segments

$x_m = 0$ → segment m is deleted

$x_m = 1$ → segment m exists

Interpret the model

$y = w_1x_1 + \cdots + w_mx_m + \cdots + w_Mx_M$

If $w_m \approx 0$ → segment m is not related to the class

If w_m is positive → segment m indicates the image is the class

if w_m is negative → segment m indicates the image is not the class

d) Decision Tree

i) Complexity of Decision Tree

$O(T_\theta)$: how complex T_θ is e.g. average depth of T_θ

We don't want the tree to be too large, thus small $O(T_\theta)$

ii) Tree regularization

Train a network that is easy to be interpreted by decision Tree

$\theta^* = argminL(\theta) + \lambda O(T_\theta)$

The objective function with tree regularization is not differentiable

Solution : https://arxiv.org/pdf/1711.06178.pdf