*Statistical Learning* May 11, 2019
Sharif University of Technology CE 40-956
H.R. Rabiee

Name: Std. Number:

# Quiz 7 (interpretable Learning)

**Questions**

1. We have three interpretation algorithms and we want to evaluate their performance on a trained model for a classification task. The model is trained on a labeled image dataset to classify images into two categories (Each image belongs to exactly one of these categories.). Each interpretation algorithm gives an importance score, for each category, to pixels of all images. So, for the $a$th image, we have two score maps $S_a^1$ and $S_a^2$ that show the importance of each pixel of this image for category 1 and 2 respectively. Design a machine based quantitative measure (which reports a number) for evaluating these three interpretation algorithms. You can use and run the trained classifier.

2. Existing approaches in interpretation can be categorized into 2 categories: (1) Perturbation and forward propagation based methods. (2) Backpropagation based methods

   (a) Explain the general idea behind each category and discuss about their advantages and disadvantages.

   (b) (The saturation problem) Why can't simple perturbation and backpropagation based algorithms interpret the following model successfully?

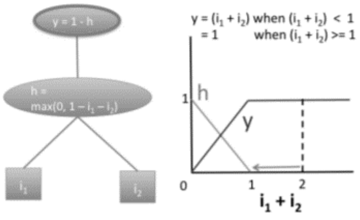

Figure 1:

   (c) To solve the above problem, instead of considering the derivation in backpropagation based models we can use the difference from a reference value. Imagine the target node t in a deep neural network. We define the $\Delta t = t - t_0$ in which $t_0$ is a reference value for example zero. We also define $\Delta x_i$ for all other nodes $x_i$ of the network. We define $C_{\Delta x_i, \Delta t}$ which shows the amount of $\Delta t$ which is triggered by the $\Delta x_i$. So:

$$\Sigma_{i=1}^{N} C_{\Delta x_i, \Delta t} = \Delta t \tag{1}$$

We also define a multiplier $m_{\Delta x_i, \Delta t}$ as:

$$m_{\Delta x_i, \Delta t} = \frac{C_{\Delta x_i, \Delta t}}{\Delta x_i} \tag{2}$$

If $y_j$s are a set of intermediate nodes between some nodes $x_i$s and $t$, show that the chain rule (3) assumption is compatible with the equation (1). In fact given $C_{\Delta x_i, \Delta y_j}$ and $C_{\Delta y_j, \Delta t}$ both satisfy (1), you should show that defining $C_{\Delta x_i, \Delta t}$ according to the chain rule also satisfies (1).

$$m_{\Delta x_i, \Delta t} = \Sigma_j \; m_{\Delta x_i, \Delta y_j} \times m_{\Delta y_j, \Delta t} \tag{3}$$

(d) Explain how this new backpropagation technique can solve the saturation problem (part b).