National University of Singapore
School of Computing
CS3244: Machine Learning
Week - 11 - Tutorial 09

RNNs and Explainable AI(XAI)

1. **RNN and BPTT** Here, we'll be computing gradients via *Backpropagation Through Time* (BPTT). The Forward Pass of a RNN can be characterised as follows(Here σ denotes *softmax* function):

$$\mathbf{h}_t = g^{[h]}((\mathbf{W}^{[xh]})^\top \mathbf{x}_t + (\mathbf{W}^{[hh]})^\top \mathbf{h}_{t-1}) \tag{1}$$

$$\hat{\mathbf{y}}_t = g^{[y]}((\mathbf{W}^{[hy]})^\top \mathbf{h}_t) \tag{2}$$

$$\hat{\mathbf{o}}_t = \sigma(\hat{\mathbf{y}}_t) \tag{3}$$

The loss L is the Cross Entropy Loss:

$$L = -\sum_t^T \mathbf{y}_t \cdot \log(\hat{\mathbf{o}}_t) \tag{4}$$

For simplicity, let's call the final time step loss, $E_T = -\mathbf{y}_T \log(\hat{\mathbf{o}}_T)$. The objective of BPTT is to update the parameters $\mathbf{W}^{[xh]}$, $\mathbf{W}^{[hh]}$, and $\mathbf{W}^{[hy]}$.

   (a) Use Chain Rule to find an expression for $\frac{\partial E_T}{\partial \mathbf{W}^{[hh]}}$. (Note, there is no need to expand the term $\frac{\partial \mathbf{h}_{T-1}}{\partial \mathbf{W}^{[hh]}}$ further)

   (b) Out of the various terms above, which one do you think is responsible for the *Vanishing Gradient Problem*?

2. **LIME**

   (a) In LIME, we use a predictor which is a quadratic function rather than a line. How do you sample points for the decision boundary? Show how quadratic LIME can improve over a linear LIME.

   (b) In LIME, an important step is to create locally perturbed data to train our surrogate model. **How to create the perturbed text data?** Suppose you are attempting a text classification task to decide whether a comment is a spam (class 1) or normal (class 0). You have trained a Transformer (a deep learning model) as a prediction model. There are two training samples: a normal comment "It looks very cute" and a spam comment "For Christmas Song visit our channel!". How to explain "For Christmas Song visit our channel!" is classified as spam by a trained Transformer?

   (c) Following the last question, **how to create the perturbed image data?** Suppose you are attempting an image classification problem. You have trained a ResNet-50 (a deep learning model) as a prediction model. Given an image of a dog, how to explain why the ResNet-50 makes such a prediction?

   (d) What are the disadvantages of using LIME for explanation?