题目：Attention Factorization machines: learning the weight of feature interactions via attention networks

Author: Jun xiao

Abstract: AFM focus on learning the importance of each feature interaction from data via a neural attention network

Introduction:

FM parameterize the weight of cross feature as the inner product of the embedding vectors of the constituent features. By learning an embedding vectors of the constituent feature, FM can estimate the weight of any cross feature.

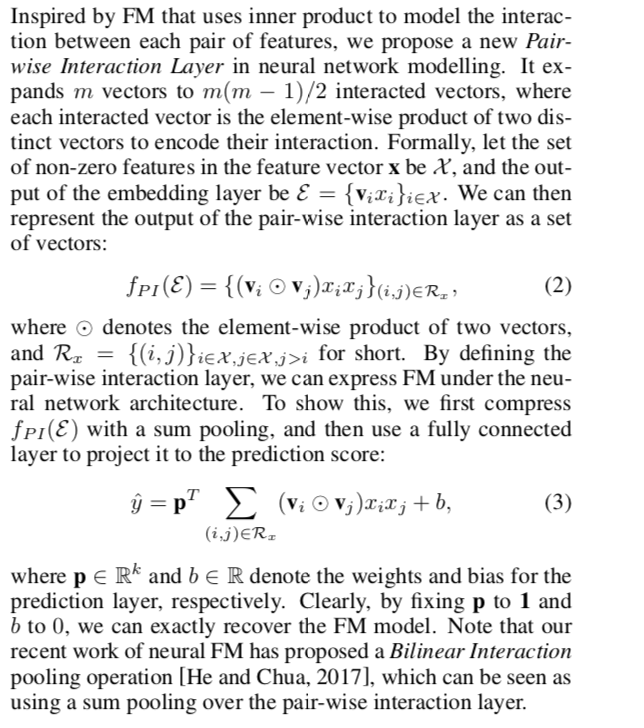
AFM is not only lead to a better performance, but also provide insight into which feature contribute more to the prediction.

FM：  
As for feature I, it share the same latent vector vi in estimating all feature interactions that the I-th feature involves. What’s more in the fm model, all estimated feature interactions wij have a uniform weight of 1. These two characteristics make the FM is not suitable for the real world.

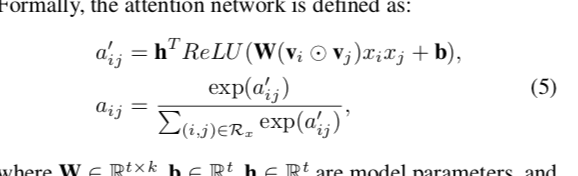
AFM:

It can be divided into five layers:

1. The sparse input: it is the same to the FM, just input the feature represent of a item
2. The embedding layer: for each feature, it will embedding a latent vector named vi, that is also the same to the FM.
3. Pair-wise interaction layer: only calculate the non-zero feature and their latent vectors’s multiply



1. Attention-based pooling layer:
   1. Use the attention mechanism to calculate the weight of different non-zero feature pair. The attention mechanism allows different parts contribute differently when compressing them to a single representation.
   2. So it problem change to how to estimate the aij, which is the parameter that evaluate the importance of different parts. We can use the minimum loss technology to calculate the aij, but this method is not suitable for the features that never co-occurred. So the author use the MLP to fix this problem.
   3. Use the attention net to train the aij



1. Learning
   1. Use the squared loss to define the loss, and use the SGD to optimize the objective function
   2. To overcome the overfitting, the author use the dropout and L2 regularization to fix out. The dropout method is used on the pair-wise interaction layer. The effect is that it can drop the useless interactions between features and improve the performance by averaging with a smaller neural networks. In the loss function we add the L2 loss. The Attention network not use the dropout because the author find out that the joint use of dropout on both the interactions layer and attention network leads to some stability issue and degrades the performance.
2. Because the weight sum at the end, AFM don’t show the power of the DNN.