推荐系统实践

林秋强

1 模型简介

1.1 FM

因子分解机(Factorization Machines, FMs)[1]在预测时，除了利用单个特征进行预测(类似于线性回归)，还利用了多个特征的复合特征(类似于使用的核技巧的SVM)，并且FM计算量比SVM更小且能处理稀疏情况。FM给每个特征都对应了一个因子，复合特征(若干的特征的乘积)的系数是对应特征的因子的积。二维情况下的FM预测结果为：

$\hat{y}(x) = \omega_0 + \sum_{i=1}^{n} \omega_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} (v_i, v_j) x_i x_j$

其中ω_0是全局偏置，ω_i是特征x_i对应的系数，v_i是特征x_i对应的隐向量。

1.2 DeepFM

DeepFM[3]是Wide & Deep Learning[2]的一种改进。它也将一个宽度模型和深度模型结合，然而宽度模型不再使用简单的广义线性模型，而是具有较好效果的FM。另一个改进是DeepFM不再给深度和宽度模型分别提供输入，而是将同一个输入经过编码后同时提供给模型的两个部分。这样的设计使得整个模型能够自动提取特征而不再需要人工挑选，并且根据文章的实验结果，还能够提升模型的准确性。模型的输出为：

$P(Y = 1|x) = \sigma(y_{FM} + y_{deep})$

其中σ是sigmoid函数，y_{FM}是宽度部分即FM的输出(具体公式见上文)，y_{deep}是深度部分即DNN的输出。

2 实验结果

使用python3.6上基于TensorFlow的DeepFM实现。DeepFM模型部分的

代码来自于[4]。

由于源文件数据量很大，为了提高实验效率，测试模型和调整参数时只使用前100万条数据作为训练集，第100万至150万条数据作为测试集。

2.1 输入特征选择

首先对数值型特征time, duration_time进行归一化处理。再对数据中的类别特征(categorical feature)进行统计，各特征中包含的类别数如表1所示。

特征名称	uid	user_city	item_id	author_id	item_city	channel	like	music_id	device
类别数	22463	373	393095	146759	386	5	2	22504	22584

表 1: 各特征中包含的类别数

可以看到uid, item_id, author_id, music_id, device中所包含的类别非常多(下文中称其为id类特征)，使用onehot变换之后输入的维数和稀疏程度都会非常大，给模型的训练造成困难。图1展示了使用id类特征的DeepFM训练时在训练集与在验证集上的精度差别(作为对比，不包含id类特征时的情况见图5)，图2展示了包含和不包含id类特征两种情况的运行时间差别。

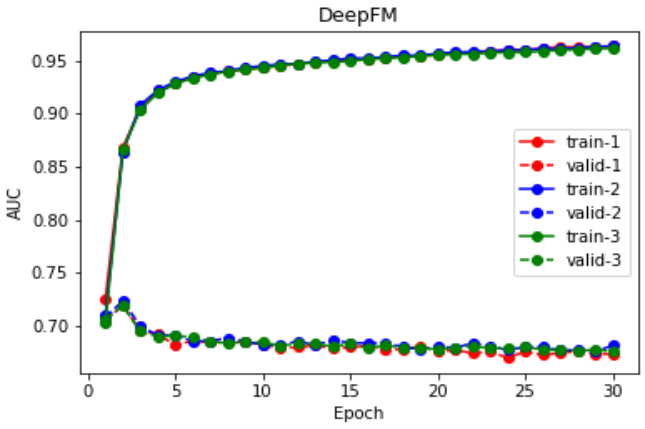

图 1: 包含id类特征时DeepFM在训练集和测试集上的AUC

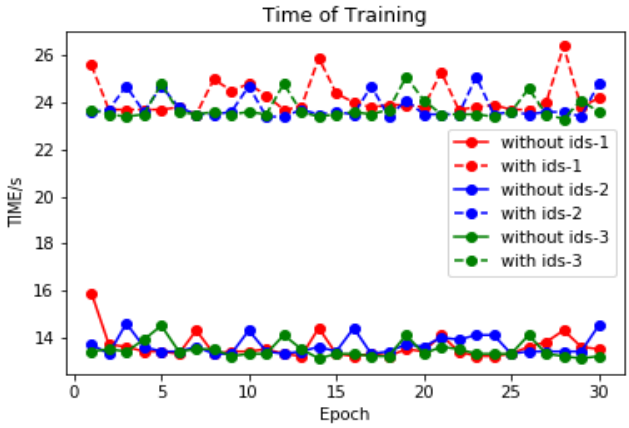

图 2: 包含和不包含id类特征的训练时间

使用id类特征时DeepFM在验证集上的平均AUC为0.67692，标准差为0.00347。在测试集上的AUC为0.70798。

可以看到使用全部的特征可以提升模型的精度，但是极大的增加了模型的训练时间，并且对训练集产生了严重的过拟合。这可能是由于过于稀疏的特征使得输入之间的相似性降低，使得模型更依赖于“记忆性”。经过实验，使用较小的数据集作为训练集时运行时间和过拟合的程度都会减少，因此也可以预期随着训练集的增大和类别数的进一步增加，训练时间和过拟合程度还会进一步增长。

使用更多的稀疏特征可以使模型精确度增加，然而训练时间和过拟合程度增加；反之使用更少的稀疏特征会使模型精确度降低，然而可以减少训练时间和缓解过拟合。因此在实际使用中要注意两者的平衡。

为了更好地体现各个模型训练难度、精度、是否容易过拟合等特点并节省运行时间，在下文的实验中输入特征均不包含uid, item_id, author_id, music_id, device。

2.2 FM/DNN/DeepFM结果对比

将DeepFM中的深度部分屏蔽，就会退化成FM；将DeepFM中的宽度部分屏蔽，就会退化成DNN。因此很容易使用相同的结构和参数运行DeepFM、FM和DNN，方便对各模型进行比较。

2.2.1 FM

FM的运行结果如图3所示。

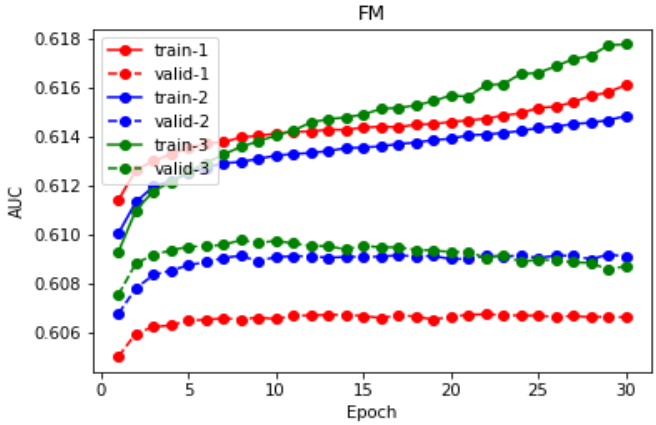

图 3: FM在训练集和测试集上的AUC

FM在验证集上的平均AUC为0.60816，标准差为0.00109。在测试集上的AUC为0.60876。

可以看出FM在训练集上可以达到较高的AUC，然而在验证集上AUC较低。事实上，如果增加训练时的epoch，模型在训练集上的AUC仍会继续增加，在验证集上的AUC仍会继续降低。这说明即使模型采用了early stopping策略，仍然在训练集上发生了过拟合。符合[2]中的描述，即宽度模型具有较好的记忆性，然而泛化能力不强。

2.2.2 DNN

DNN的运行结果如图4所示。

DNN在验证集上的平均AUC为0.60671，标准差为0.00154。在测试集上的AUC为0.60969。

可以看出DNN在训练集和验证集上的AUC的比较相近，但是不如FM在测试集上达到的AUC。并且在验证集上的结果更不稳定。

2.2.3 DeepFM

DeepFM的运行结果如图5所示。

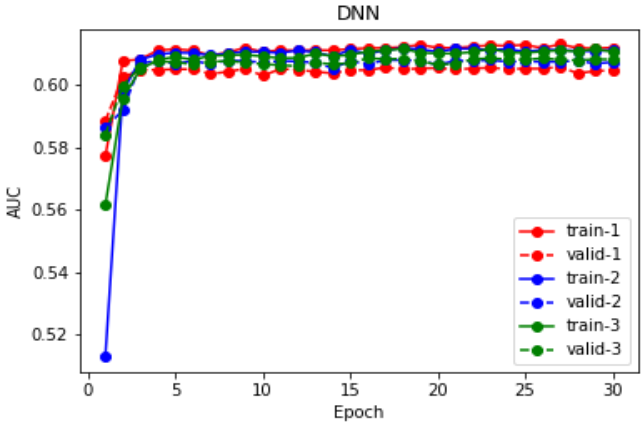

图 4: DNN在训练集和测试集上的AUC

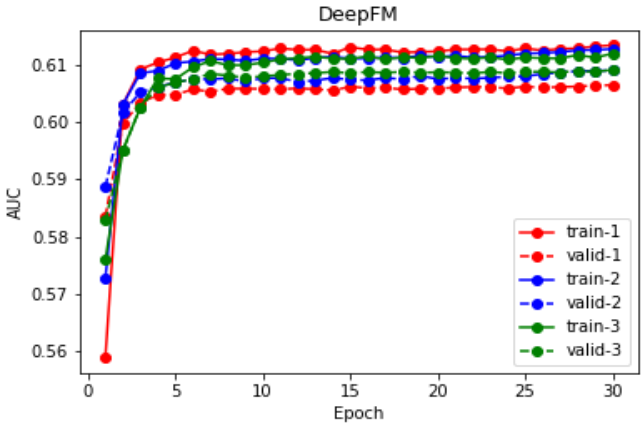

图 5: DeepFM在训练集和测试集上的AUC

DeepFM在验证集上的平均AUC为0.60817，标准差为0.00124。在测试集上的AUC为0.60989。

可以看出DeepFM在训练集上的效果比FM略差，比DNN更好；而在验证集上的效果比后两者都要更好。说明DeepFM结合了FM和DNN的优点，一方面对训练集中数据具有较好的记忆性，在训练集上达到较好的预测效果；另一方面也具有较好的泛化能力，在验证集上的预测效果和在训练集上相近并且更加稳定。

2.3 结论

从实验中，我们可以得出以下结论：

1. 数值型的特征需要进行归一化或者标准化。如果不进行归一化或标准化，并且输入的量级差别很大时，模型将会无法进行学习。例如本实验中如果不将time归一化，训练将无法进行，而不将duration_time归一化对模型的影响较小；

2.类别型的特征如果包含的类别数量很多，要谨慎考虑它对模型的作用。对类别特征进行onehot变换后可能会产生非常巨大的维数，一方面增加了模型的复杂度，增大了训练所需时间；另一方面过多的参数和过于稀疏的输入可能使得模型更容易过拟合乃至训练失败。因此遇到包含的类别数量特别多的特征时，应该权衡精度和时间/空间效率，必要时考虑对该特征进行进一步处理乃至直接舍弃该特征；

3. 宽度模型对训练数据有较好的的记忆性，但是泛化能力不足。随着训练的进行，宽度模型(本实验中为FM)在训练集上的预测精度逐渐增加，然而在验证集上的精度反而逐渐下降，即使使用了early stopping等防止过拟合的技巧依然如此；

4. 深度模型具有较好的泛化能力，在训练集和验证集上的预测精度比较接近，然而在训练集上的精度不如宽度模型。

5. 将宽度模型和深度模型结合可以结合两者的优点，具有较好的记忆性的同时也具有较好的泛化能力，并且预测精度也比宽度模型和深度模型都要高。

参考文献

[1] Rendle S. Factorization Machines[C]// IEEE International Conference on Data Mining. 2011.

[2] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems[C]//Proceedings of the 1st workshop on deep learning for recommender systems. ACM, 2016: 7-10.

[3] Guo H, Tang R, Ye Y, et al. DeepFM: a factorization-machine based neural network for CTR prediction[J]. arXiv preprint arXiv:1703.04247, 2017.

[4] https://github.com/ChenglongChen/tensorflow-DeepFM/blob/master/DeepFM.py