Wide & Deep Learning for Recommender Systems

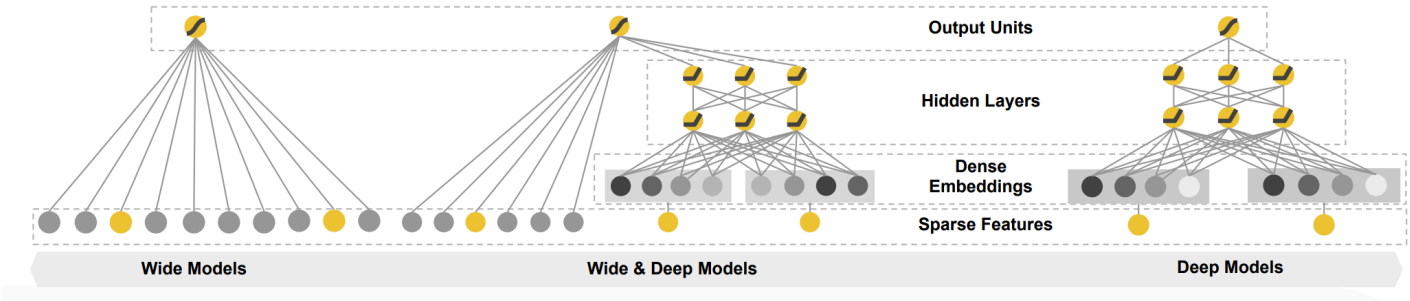
Introduction

推荐系统可看作一个检索排序系统，输入是用户和上下文信息，输出是物品的排序列表。同检索排序系统一样，推荐系统的难点是同时实现 *memorization* 和 *generalization*，两者的定义为：

1. memorization: 从历史数据中发现 item 或特征之间的相关性；
    - 之前大规模稀疏输入的处理是：通过线性模型+特征交叉，通过特征交叉能够带来很好的效果并且可解释性强，但是 generalization(泛化能力)需要更多的人工特征工程。

2. generalization：相关性的传递，发现在历史数据中很少或者没有出现的新的特征组合。
    - 相比之下，dnn 几乎不需要做特征工程，通过对低维 dense embedding 进行组合可以学到更深层次的隐藏特征。但是，缺点是会导致 over-generalize (过度泛化)。推荐系统中变现为：会给用户推荐不是那么相关的物品，尤其是 user-item 矩阵稀疏并且 high-rank 时。

Wide & Deep 模型通过联合训练一个线性模型部分和一个神经网络部分，可以在一个模型时同时实现 memorization 和 generalization。



Wide 部分

wide 部分是一个线性模型 $y = w^T x + b$，如上图左所示，其中 y 为预测结果，$x = [x_1, x_2, \cdots, x_d]$ 为 d 维向量，$w = [w_1, w_2, \cdots, w_d]$ 为模型参数，b 为 bias。特征集合包括原始输入特征和转换特征。一个重要的转换为交叉积转换，定义为：

$$\phi_k(x) = \prod_{i=1}^{d} x_i^{c_{ki}} \quad c_{ki} \in \{0, 1\}$$

其中，c_{ki} 是布尔变量，当第 i 个特征是第 k 个转换 ϕ_k 的一部分时，该值为1，否则为0，对于 binary 特征，交叉积转换 (比如 "AND(gender=female,language=en)") 当且仅当包含的特征 ("gender=female" 和 "language=en") 都为 1 时才为 1，否则值为 0。这提取了 binary 特征的组合信息，在线性模型中加入了非线性元素。

Deep 部分

deep 部分是个前向神经网络，如上图右所示。每个稀疏、高维 categorical feature 转换为低维、稠密的实数向量，称为 embedding 向量。embedding 向量随机初始化，然后经过训练使损失函数达到最小值。这些低维的稠密 embedding 向量输入至神经网络的隐藏层：

$$a^{(l+1)} = f(W^{(l)}a^{(l)} + b^{(l)})$$

Joint Training

模型的联合训练使用 mini-batch stochastic optimization 反向传播梯度，由输出层到 wide 和deep 部分同时进行。在试验中，wide 部分使用 Follow-the-regularized-leader (FTRL) 算法，并使用 L_1 正则，deep 部分使用 AdaGrad。

对 logistic regression 问题，模型的预测结果为：

$$P(Y = 1|x) = \sigma(w_{wide}^T[x, \phi(x)] + w_{deep}^T a^{(l_f)} + b)$$

模型的结构如下图所示：