# Wide&Deep算法设计文档

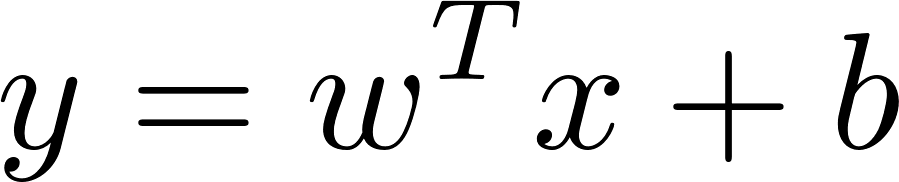
小组成员：王胜广、龚润宇、梅楚鹤、柳俊志、尹国健

## 1. BPR算法介绍

### 1.1 背景

Wide and deep 模型是 TensorFlow 在 2016 年 6 月左右发布的一类用于分类和回归的模型，并应用到了 Google Play 的应用推荐中 [1]。wide and deep 模型的核心思想是结合线性模型的记忆能力（memorization）和 DNN 模型的泛化能力（generalization），在训练过程中同时优化 2 个模型的参数，从而达到整体模型的预测能力最优。

### 1.2建模

(1) Wide侧用的模型是逻辑回归(logistic regression, LR)。属于广义线性模型的范畴。

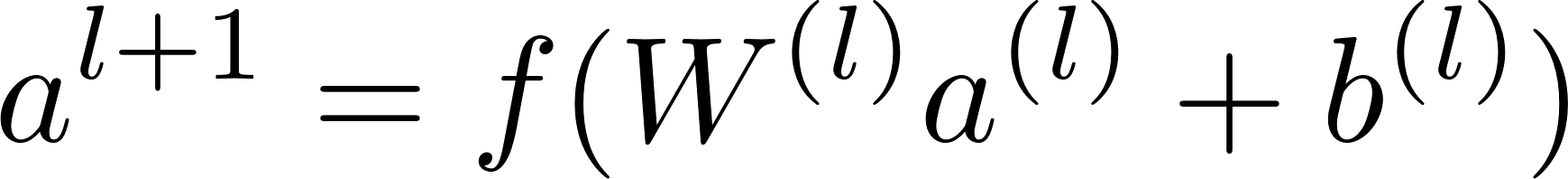
- 输入：离散特征（连续特征离散化）+人工交叉特征

- 记住那些稀疏的，特定的rules

- 优化器：FTRL

- 特征交叉：”性别=女 and 语言=英语”，只有二者同时成立，交叉特征的值才为1。限制就是不能生成从未在训练数据中出现过的query-item特征对。

(2) Deep就是指 Deep Neural Network



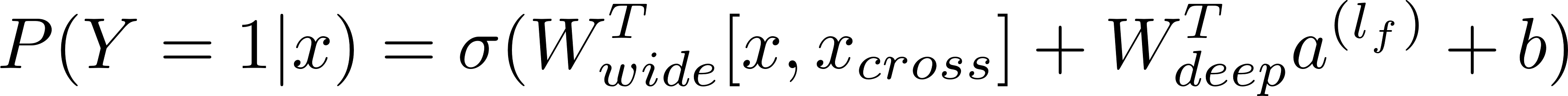
- 输入：连续特征+离散特征embedding

- 泛化推荐一些相似的items

- 隐层激活函数：ReLU，修正线性单元

- 优化器：AdaGrad

(3) Wide & Deep：线性模型 + DNN部分



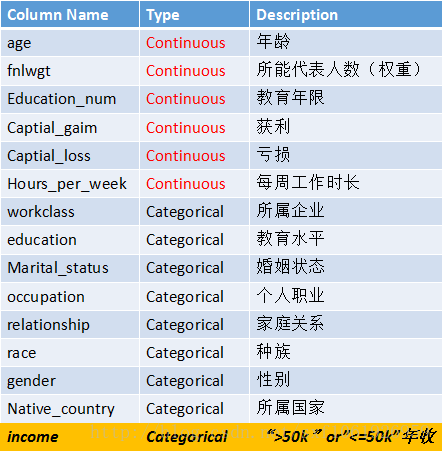
- 激活函数： sigmoid

- 损失函数：common logistic loss function

- 反向传播，loss function对w求偏导后，两侧参数的学习过程与单wide或单deep模型时保持一致

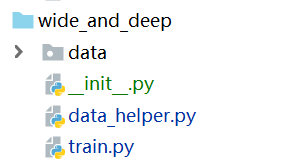
## 2 编程实现

数据集：人口普查数据集（<https://archive.ics.uci.edu/ml/datasets/adult>）。包含48842个人口普查信息。用来预测收入是否超过五万美元。



### 2.2 代码详解

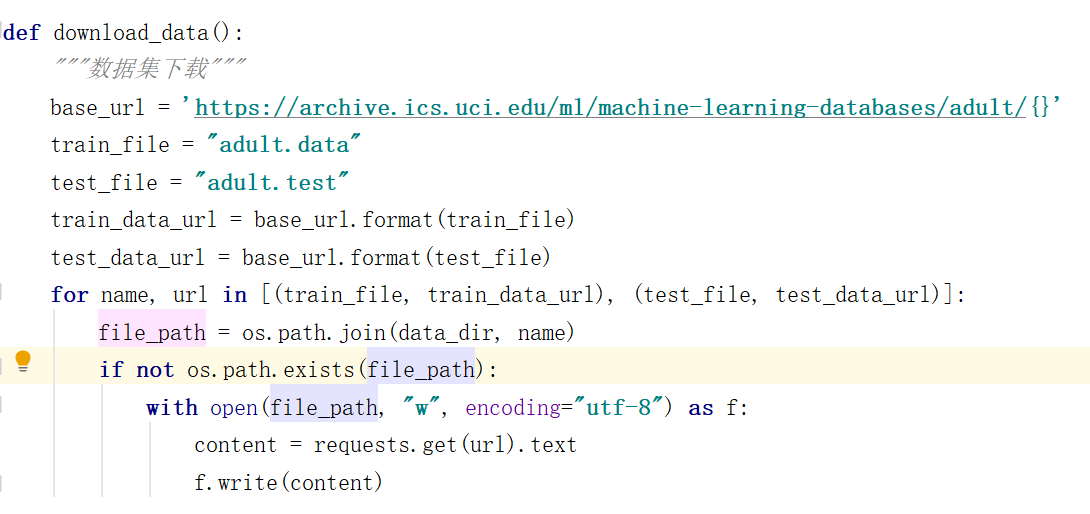
代码data\_helper.py和train.py两个文件，data\_helper.py为数据集下载载入，数据集切分等，train.py为模型定义训练等。



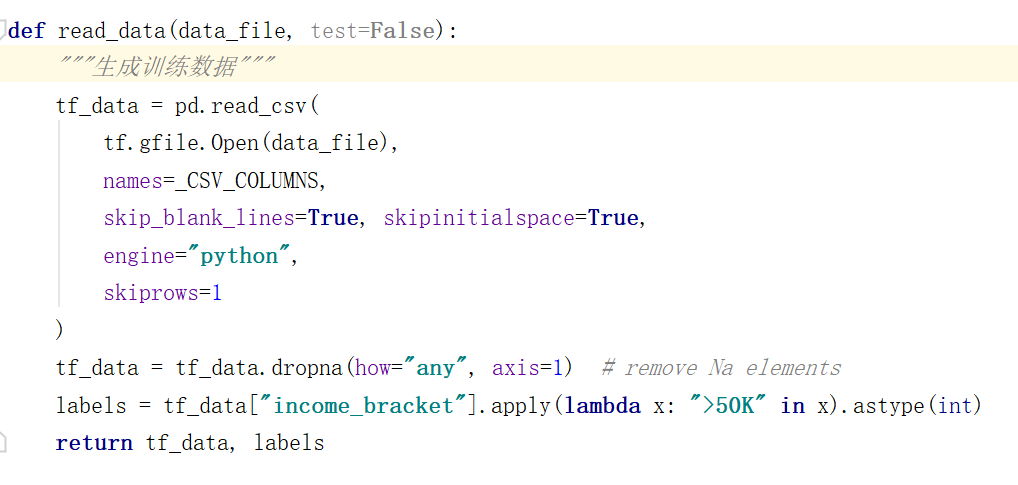
#### 2.2.1 数据载入（data\_helper.py）

**1）数据集下载切分**

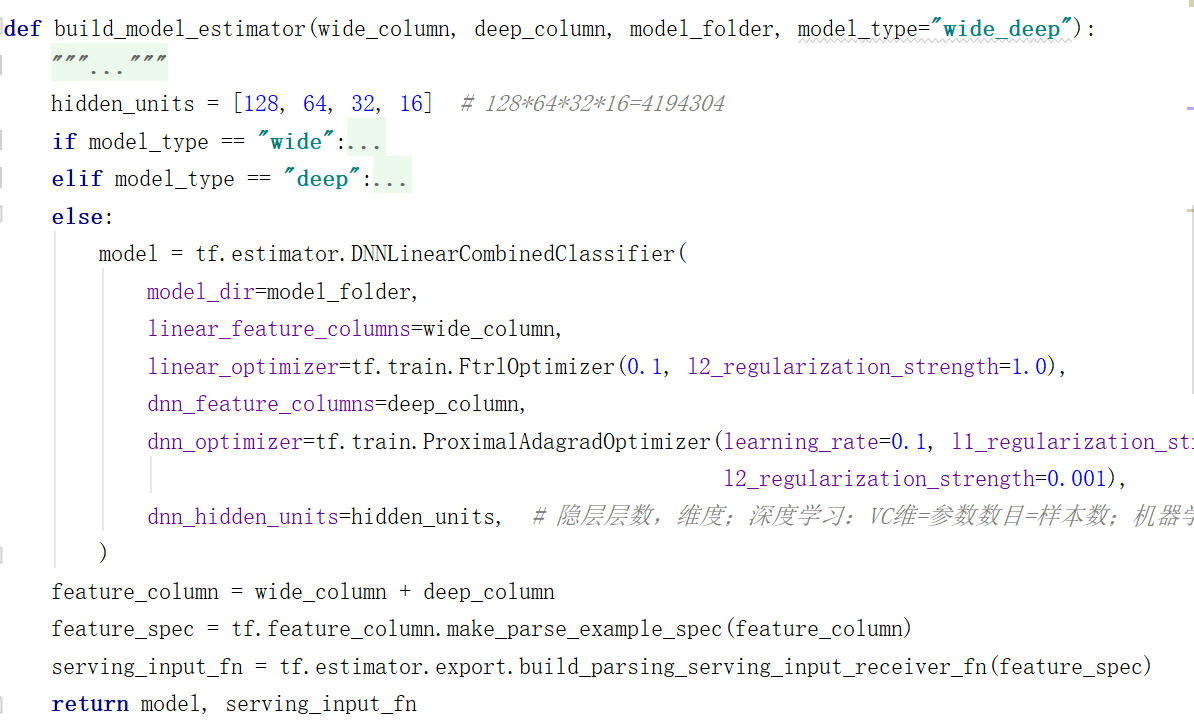
数据集已经分为训练集和测试机，故无需自己切分



**2）生成训练数据**

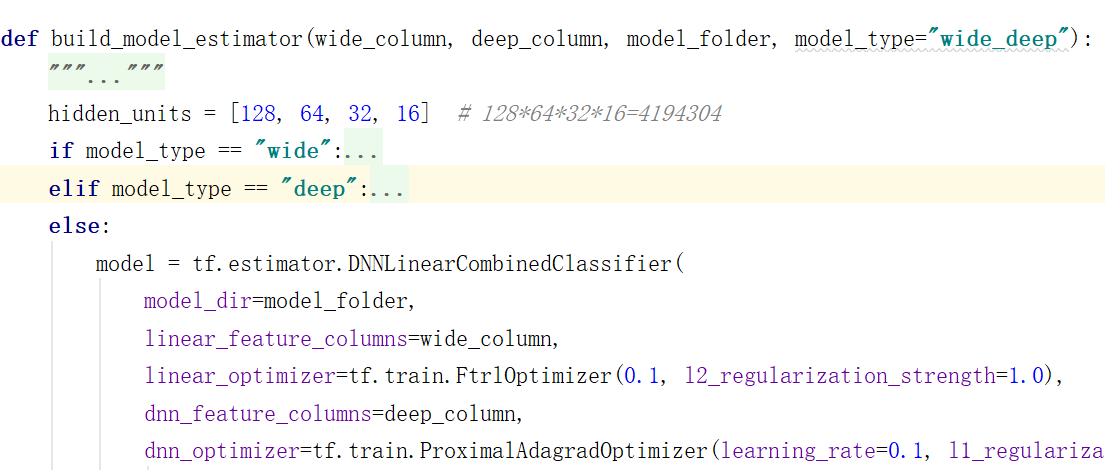


#### 2.2.2 模型构建



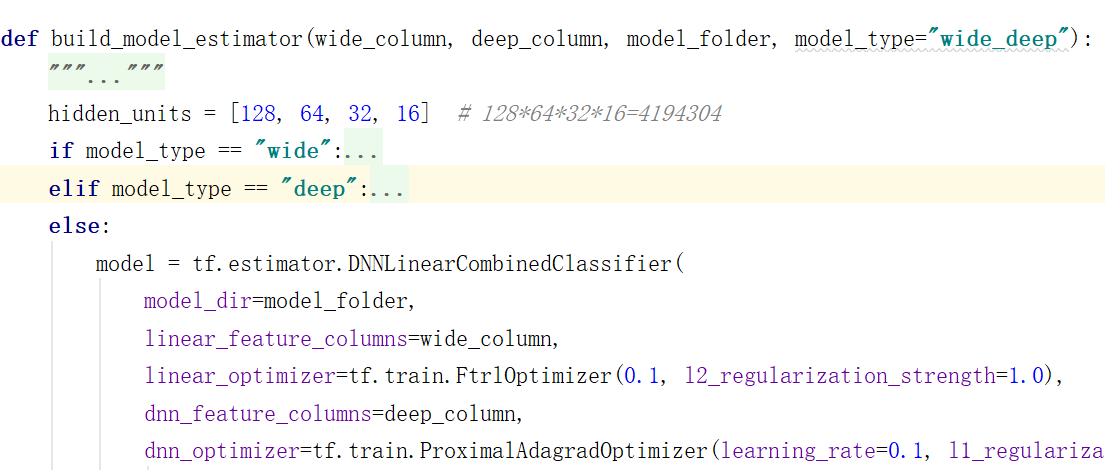
#### 2.2.3 模型保存

模型构建中使用tensorflow的高阶api，训练过程中会自动保存模型；

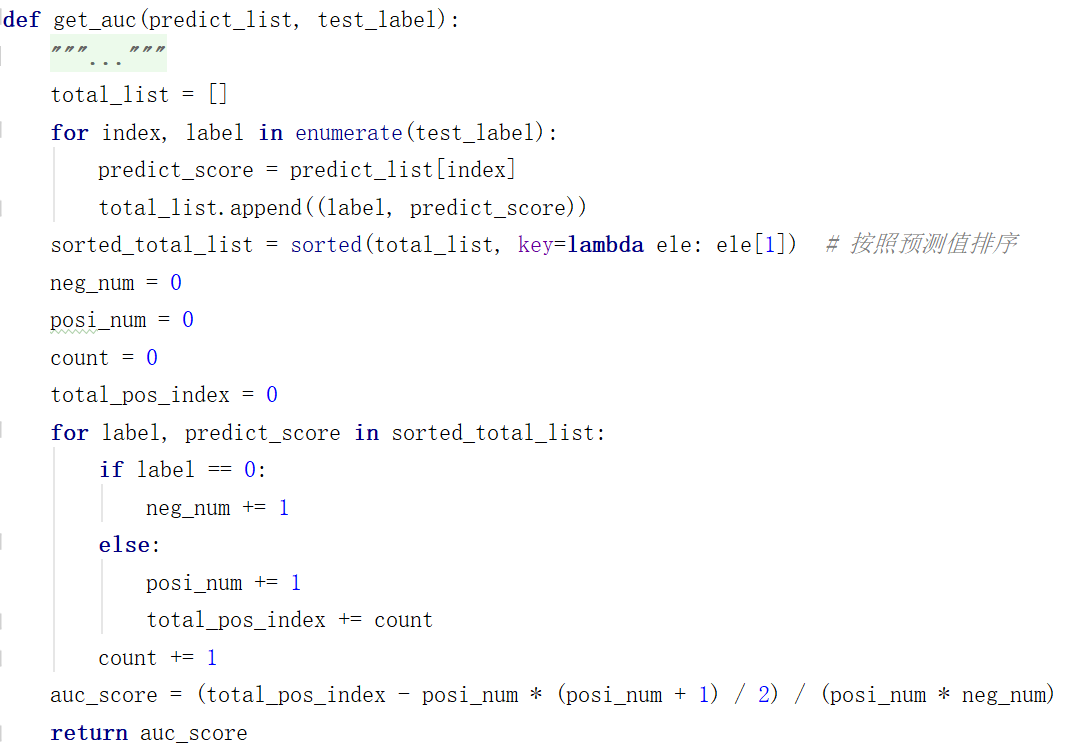


#### 2.2.4 模型加载

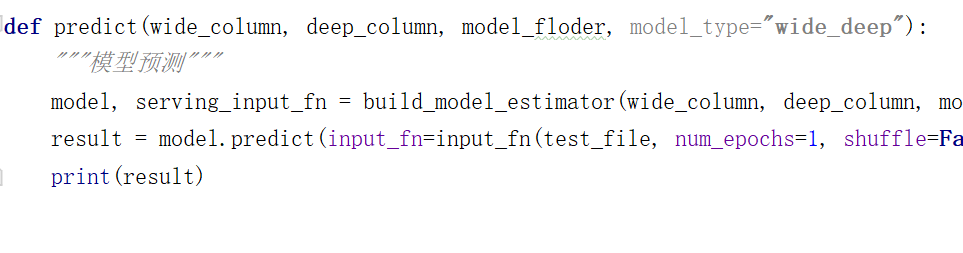
模型构建中使用tensorflow的高阶api，直接调用build\_model\_estimator,传入对应model\_dir，会载入上一次保存的模型。



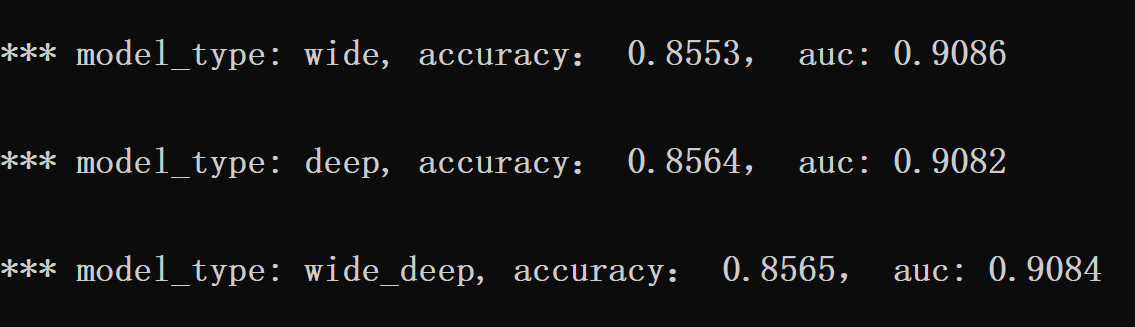
#### 2.2.5 AUC



#### 2.2.6 推荐预测



### 2.3 运行截图



## 3 总结

* 本文提出了新的模型结构，实现了对memorization和generalization的统一建模。
* 缺点：Wide部分还是需要人为的特征工程。
* 改进：DeepFM，FM代替Wide部分，FM模块实现了对于1阶和2阶组合特征的建模，不再需要人工特征。