炼丹调参

随机数种子

kaggle上流传着一个传说，随机数种子是最关键的超参数【笑】，据统计，比较流行的随机数种子是42【银河漫游指南粉丝快点赞】，2021【年份】，1234等。那为什么随机数种子这么重要呢- -。

因为机器学习&深度学习模型参数初始化、数据shuffle、数据batch生成、特征抽样等环节有大量随机性存在，因此为了能够规避随机数所带来的影响，聚焦模型超参数改变本身带来的性能提升，通常我们需要固定随机数种子。

这里给出一个案例：

```
import os
import random
import torch
import numpy as np


def seed_everything(seed=1234):
    random.seed(seed)
    os.environ['PYTHONHASHSEED'] = str(seed)
    np.random.seed(seed)
    torch.manual_seed(seed)
    torch.cuda.manual_seed(seed)
    torch.backends.cudnn.deterministic = True

seed_everything()
```

其中在使用sklearn.model_selection.GridSearchCV等方法时，还需要自己手动初始化CrossValidation方法，传入cv参数中，否则依然无法保证数据划分的一致性。

网格搜索

暴力搜索的近似，将搜索空间离散化，分辨率越高精度越高，但速度越慢。优点：确定性，全局性。缺点：低效。**适合参数较少的场景，例如SVM（C、kernel、gamma）。**

- 网格搜索使用参考
- 工具：from sklearn.model_selection import GridSearchCV
- 代码案例： https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_digits.html#sphx-glr-auto-examples-model-selection-plot-grid-search-digits-py

随机优化

随机搜索提供了一种更高效的解决方法（特别是参数数量多的情况下），Randomized Search为每个参数定义了一个分布函数并在该空间中采样（sampling），论文Random search for hyper-parameter optimization进行了分析和实验。

- 工具：from sklearn.model_selection import RandomizedSearch

- 代码案例：用法完全类似网格搜索，但需要定义采用概率分布。 https://scikit-learn.org/stable/auto_examples/model_selection/plot_randomized_search.html#sphx-glr-auto-examples-model-selection-plot-randomized-search-py

贝叶斯优化

- 贝叶斯优化直觉理解
- [拓展]贝叶斯优化/Bayesian Optimization

遗传算法

- 遗传算法的直觉理解
- 遗传算法案例：Kaggle-moa竞赛，遗传算法优化KNN 权重

粒子群算法

- 粒子群算法直觉理解
- 遗传算法可视化
- [拓展]遗传画师

Optuna

强烈推荐，深度学习&机器学习量身定制的参数优化库

- 项目地址： https://github.com/optuna/optuna
- 中文介绍： https://www.zhihu.com/question/384519338/answer/1206812752

项目

在titanic项目中使用optuna库进行参数优化，提升模型效果。

Refs

- Bergstra J, Bengio Y. Random search for hyper-parameter optimization[M]. JMLR.org, 2012.