极客大学

极客大学机器学习训练营 Python/R 中的数据预处理和可视化

王然

众微科技 AI Lab 负责人

二〇二一年一月四日

大纲

极客大学

大纲

极客大学

内容简介

本章主要内容是讲解常见的数据处理方法。大部分内容在 Jupyter Notebook 当中。虽然如此，本章的目的是为了之后的学习做准备。所以增加了一些内容。

工具和工具简介

极客大学

- numpy：矩阵（tensor）处理工具。
- Jax：打了鸡血的 numpy。
- Pandas：一般数据处理工具。
- dplyr：探索性数据分析很好的工具。
- Matplotlib：探索性数据分析的画图工具。
- TensorBoard：记录模型训练过程的工具。

本章难点

极客大学

- numpy 当中的 einsum 和 broadcast 转换。
- Jax 函数式编程的习惯。

大纲

极客大学

详细内容见 notebook。

- 核心加速原理：JIT+Jaxpr+Asyn Dispatch+XLA
- 核心写作限制：Tracable 和 Pure 函数

Jax 核心之 JIT 和 Jaxpr

极客大学

- JIT = Just-in-time Compilation.
- 第一次运行时候进行编译：充分利用各种 runtime information
- 可以指定哪些可以重新编译（static_argnums），这个也会解决 trace 的问题；我们在这里不会对 trace 问题和 Jax Primitives 做介绍（后面讲到时候再介绍）

Async Dispatch 和 pure 函数

极客大学

- 还记得上次说的异步图的事情了么？Jax 可以自动实现。
- 原因：函数没有负效应。
- 问题：需要改写一些常见用法。见 notebook。

XLA

- XLA = Accelerated Linear Algebra. 详见XLA 官方文档
- 这保证 Jax 可以在各种 device（CPU/GPU/TPU）运行。
- 也可以自定义 XLA 算子，如果觉得本身效率不够高。但一般不需要。

大纲

极客大学

见 notebook。

大纲

极客大学

见 notebook。

大纲

极客大学

总结

极客大学

- 本章大部分内容只需要用的时候速查即可；
- einsum 和 broadcast 如果不会没有关系，我们将会在 DL 章节重新讲解；
- 重点：玩一玩 Jax，很好玩的 ~

预习

极客大学

- 下一章对于数学基础较为缺乏的同学来说，会有一定困难。
- 预习内容和之前类似，不要要求都会，我们大部分会讲。
- 但是可以找一些看数学的感觉，比如说做一些第一章的题。