

极客大学机器学习训练营

数据预处理和可视化

王然

众微科技 AI Lab 负责人

二〇二〇年十二月二十八日



- 1 大纲
- 2 Numpy 和 Jax
- 3 Pandas 和 R
- 4 Matplotlib 和 TensorBoard
- 5 要点复习
- 6 参考文献



- 1 大纲
- 2 Numpy 和 Jax
- 3 Pandas 和 R
- 4 Matplotlib 和 TensorBoard
- 5 要点复习
- 6 参考文献



- ▶ 有人说数据科学家 90% 的时间都在进行数据预处理，10% 的时间在做建模。
- ▶ 这句话是正确的，但是需要正确的进行解释。
- ▶ 数据预处理包含两种可能性：
 - ▶ 取数。这种往往是根据规则从数据库当中将数据取出来，这类工作严格来说是 ETL 工程师来做的（虽然很多时候数据科学家可能要帮助 ETL）→ 关于这一点请注意，一定要清楚的跟 ETL 进行沟通，反复的取数是浪费时间的罪魁祸首之一。
 - ▶ 数据处理。这部分数据处理，是数据已经取出来之后要做的。
 - ▶ 上线的处理。上线后，遇到异常数据处理。

为什么不让 ETL 把所有事情都干了

- ▶ 建模过程中，有些数据处理是和模型方法论有关系的。
- ▶ 上线中，很多异常只有数据科学家才知道。

- ▶ SQL → 难以自定义操作
- ▶ Spark → 自定义操作有时更麻烦，对于用来测试的中等数据集速度不如内存式。
- ▶ numpy → 矩阵和张量的处理。
- ▶ Pandas 和 R 的 dataframe。两个来源一样，是我们主要的工具。
- ▶ dplyr → 一种 R 当中的语法，很方便进行探索。



- ▶ 目的不是为了画出漂亮的图写报告。
- ▶ 目的是提升模型预测的效果。
- ▶ Matplot 只会介绍对探索性分析最有用的画图方法。
- ▶ TensorBoard 只会简单介绍在 DeepLearning 当中使用。



那么关于 Jax 是干嘛的呢？

- ▶ 打了鸡血的 numpy。
- ▶ 实现 async 的数据处理。
- ▶ 加了 Jit 的优化器。
- ▶ 函数式编程的一次预习。



- 1 大纲
- 2 Numpy 和 Jax
- 3 Pandas 和 R
- 4 Matplotlib 和 TensorBoard
- 5 要点复习
- 6 参考文献



- ▶ 主要内容见 jupyter notebook。
- ▶ 重点：Broadcast, einsum。



- ▶ Jax 的核心是对 Numpy 进行优化和实现 AutoGrad。更主要的设计是设计出类似于 TensorFlow 一样（甚至超过的效率），但是减少不熟悉的界面。
- ▶ Jax 的四层优化策略：代码写法 → Jaxpr → JIT → XLA



- ▶ 代码要求必须用 pure 函数（没有副效用）。
- ▶ 效果：可以实现自动的并行。比如说我输入 a, b, c, d , $x = a + b, y = c \times d$, 而最终结果要的是 $x \times y$, 则计算 x, y 时候自然可以同时并行。这是因为计算 $a + b$ 的时候不会改变 c 和 d 的值。
- ▶ 所以自身是 Jax 自身是 Async 的。
- ▶ 当然一些东西就需要用其他的方法，例如 if else while 等。这些都有替代，详细见 notebook。

- ▶ 一种中间语言；使用中间语言的好处是优化更容易，而且你可以看到优化结果。
- ▶ 结合 JIT 用法极爽。
- ▶ 但是要生成合适的 Jaxpr（以及保证有效），必须用 `jax` 和 `flax` 内部操作。不要自己引入其他函数。
- ▶ 如果要引入，必须定义所谓的 Primitive。具体比较麻烦，我们在后面会讲。

- ▶ 在代码第一次运行时候进行编译。目的，利用编译的结果重新优化。
- ▶ 通过不同的 JIT 指令，可以查询到 Jaxpr。如果你对 Jaxpr 不满意，那么就可以修改。

- ▶ 来源于 TensorFlow 的加速方法。见[介绍](#).
- ▶ 暂时我们不会做更多讲解。我们在讲到深度学习框架的时候会进行讲解。
- ▶ 你只需要记住，通过 Jax 你可以定义自己的 XLA 算子。所以你可以对层次进行优化。

- 1 大纲
- 2 Numpy 和 Jax
- 3 Pandas 和 R
- 4 Matplotlib 和 TensorBoard
- 5 要点复习
- 6 参考文献



见 notebook。



- 1 大纲
- 2 Numpy 和 Jax
- 3 Pandas 和 R
- 4 Matplotlib 和 TensorBoard
- 5 要点复习
- 6 参考文献



见 Notebook



- 1 大纲
- 2 Numpy 和 Jax
- 3 Pandas 和 R
- 4 Matplotlib 和 TensorBoard
- 5 要点复习
- 6 参考文献



本章所讲述内容比较简单，所以大部分时候，只需要去寻找对应 API 即可。但是以下为难点：

- ▶ Broadcast 和 Einsum；尤其是两者转换。
- ▶ Jax 的函数式变成习惯。



- ▶ Python 数据分析最好的是McKinney (2012)。
- ▶ numpy, Matplotlib, TensorBoard 相对来说比较简单, 只要看官网就可以了。
- ▶ jax/flax 的资料目前只有官网, 世界上尚无比较系统的书籍。我和博文视点将会在明年出版第一版书籍, 敬请期待。





下一周的内容将会是最为 hard-core 的一周。

- ▶ 预习微积分和矩阵代数内容。
- ▶ 请阅读 All of Statistics(Wasserman 2013) 的 1, 2, 3, 6, 9, 11 章。十分重要, 看不懂地方请一定注明。



- 1 大纲
- 2 Numpy 和 Jax
- 3 Pandas 和 R
- 4 Matplotlib 和 TensorBoard
- 5 要点复习
- 6 参考文献



-  McKinney, Wes (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.* " O'Reilly Media, Inc."
-  Wasserman, Larry (2013). *All of statistics: a concise course in statistical inference.* Springer Science & Business Media.

Thanks!

