

DVC

DVC or Data Version Control is an open source tool for managing data assets. It is very useful but also can be quite overwhelming to use.

The main use cases I've found for DVC are:

1. Keeping large data assets (e.g. machine learning datasets) version controlled alongside code so that you always know where the latest version of some project specific training data can be found.
2. Making sure that all intermediate and final outputs from experiments are reproducible. This is always good to know when a client inevitably asks you "ok how did you get **Y** surprising result?" or "can you just confirm that you included **X** features in your model?"
DVC helps by:
 1. Keeping track of file hashes used at each stage in a pipeline
 2. Keeping a copy of the file content at each stage in the pipeline.

Revision #7

Created 24 March 2022 17:54:07 by James

Updated 21 January 2024 14:49:26 by James