Интеграция с DVC и S3

Общая информация

DVC представляет собой инструмент для версионирования нетекстовых данных (например, файл с моделью *.pkl или датасеты в* .pkl), дополняющий возможности Git. Также он помогает обеспечивать полноценную воспроизводимость экспериментов машинного обучения.

DVC связывает код модели и версии данных, это в свою очередь предоставляет возможность версионировать данные всевозможных типов и размеров, использовать удаленное хранилище для хранения и передачи, а также позволяет сохранять и сравнивать различные метрики моделей.

Благодаря DVC существует возможность в Git хранить данные в виде метаданных, а сами данные располагать в удаленном или локальном хранилище в другом месте.

DVC является своего рода плагином для S3 хранилища. Основным понятием в S3 является бакет (“bucket”-корзина), который можно сравнить с директорией в файловой системе. На платформе по умолчанию реализовано два бакета: первый бакет используется для хранения файлов атрибутов моделей, а второй выступает в роли хранилища для информации, которая логируется в инструменте Mlflow.

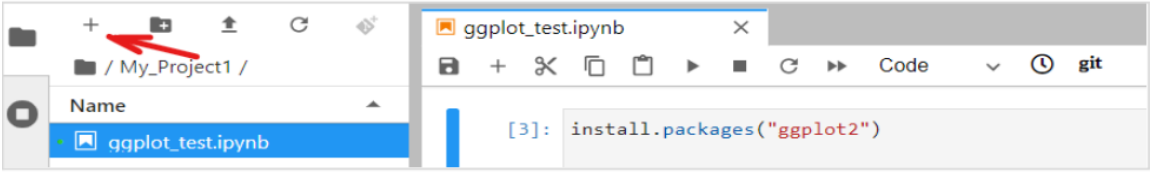
Работа с терминалом в JupyterLab

Терминал JupyterLab предоставляет возможность работы с интерпретаторами команд операционной системы, такими как bash, tcsh и т. д.

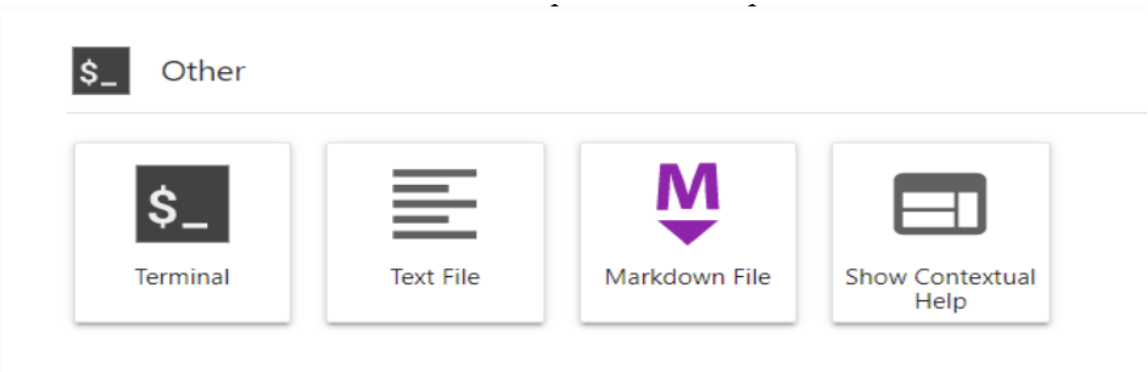
Терминалы запускаются в рамках того же инстанса, на котором запущен ваш Jupyter Server. Команды выполняются с правами вашего пользователя.

Работа с терминалом может быть полезна или даже необходима, в случаях, когда требуется установить новые библиотеки/фреймворки, при работе с системой контроля версий Git или для сбора информации при возникновении ошибок.

Для того, чтобы открыть новый терминал нажмите на символ “+” в левой панели основного интерфейса.



В появившейся вкладке **“Launcher”** выберите новый Терминал.



Примечание: подробно о работе с Терминалом Jupyter Lab можно прочитать в официальной документации:
https://jupyterlab.readthedocs.io/en/stable/user/terminal.html

Основные команды

Интеграция Git с DVC происходит на этапе работы с pkl-файлами и файлами, имеющими большой объем (архивы, датасеты, картинки). Система контроля версий Git не предназначена для хранения и версионирования бинарных (и других нетекстовых) файлов, особенно когда эти файлы большого объёма. В развернутой Системе подразумевается, что в DVC будут отслеживаться *.pkl файлы с обученной моделью (например, мы обучили модель sklearn.tree.DecisionTreeClassifier и сохранили этот объект в файл model.pkl). Опционально, можно сохранить в DVC датасеты, на которых модель обучалась, в этом случае нужно убедиться, что в удаленном S3 хранилище достаточно свободного места для записи, так как объем датасетов может быть достаточно большим.

Чтобы сохранить в хранилище S3 сериализованный файл с моделью model.pkl, который должен находиться в директории pkl, нужно выполнить следующую команду в терминале (текущей директорией терминала должен быть корень текущего проекта):
`dvc add ./pkl/model.pkl`

После выполнения команды в папке pkl помимо model.pkl появится файл model.pkl.dvc, в котором указан хэш файла, размер и название, по которым DVC сможет определить с каким файлом model.pkl связан наш коммит в Git.

Затем следует выполнить команду:
`dvc push`

Внимание! Для дальнейшего удобства работы в системе следует придерживаться подхода, при котором для каждой версии модели сначала выполняются команды `dvc add`, `dvc push`, а затем `git add -> git commit -> git push origin dev`. При таком

подходе будет легко ориентироваться в коммитах git, поскольку каждому из них будет соответствовать версия модели в DVC.

После выполнения команды файл model.pkl, будет сохранен в удаленном S3 хранилище, а Пользователь при коммите в git будет должен добавить на отслеживание (track) файл в директории pkl/model.pkl.dvc (подробнее см. п. «Взаимодействие с системой контроля версий Git»). Отметим, что в этом файле содержится md5 хэш, причем первые два символа в нем используются как имя для папки в бакете S3, в которую непосредственно отправляет сам model.pkl на хранение. Например, если в файле model.pkl.dvc строка с md5 хэшем выглядит так “md5: df175d022e2e43b425d1ea70adc439f6”, то значит файл model.pkl в S3 внутри бакета ‘b-a’ будет создана папка с названием “df”, в которую и будет загружен файл с моделью.