Машинное обучение: инструменты и технологии

MADE academy

Эмели Драль

Data analysis tools and technologies

1. Operation systems
2. Code repository
3. Software engineering
4. Administration
5. Data bases, SQL
6. ETL, Pipelines
7. Visualization, dashboarding
8. Distributed computing
9. Cloud Platforms

Operation systems

OS

OS Unix

- Terminal
- Bash
- Remote server
- Setting up, updating the system
- Virtual environment
- Package managers: apt-get, aptitude

OS

Mac OS

- Terminal
- Bash
- RDP
- Setting up, updating the system
- Virtual environment
- Development kit: xcode, homebrew

OS

OS Windows

- Terminal
- Putty
- WSL - **W**indows **S**ubsystem for **L**inux
- RDP
- Setting up, updating the system
- Virtual environment

Code repository

Code repository

Control version systems

- git, svn, cvs
- most popular: git, github, gitlab
- Setting up a repository
- Groups, members, access rights
- Cloning, push/pull
- Branches
- Code review & pull requests

Code repository

Git, github/gitlab

- Open source distributed version control system
- Public and private repositories
- Support local and remote repositories
- Render jupyter notebooks
- Online tutorial: https://try.github.io/

Software engineering

Software engineering

Software developments

- Scripting language (solid knowledge)
- Compiled language (basic understanding)
- Development environment (IDE)
- Testing approaches
- Debugging
- Code style
- Software system architecture

Software engineering

Scripting language (Python)

Python (preferably), R

- Programming paradigms
- Syntaxis
- Standard Template Library
- Libraries for Data Science
- Interactive mode, scripting mode, package mode

Software engineering

Python libraries

- Standard libraries: os, math, collections, datetime, json, etc
- Pandas, Numpy, Scipy, Scikit-learn
- Matplotlib, Seaborn, Plotly
- Python packages for popular ML tools: LightGBM, XGBoost, Catboost, Tensorflow, VowpalWabbit
- Pytorch, Keras
- Keras-RL, Openai

Software engineering

Compiled programming language

- At least read C++/Java code

Data Bases

Data Bases

Data sources

- File systems
- SQL DB
- noSQL DB

Data Bases

SQL & noSQL DB

- Access rights
- Reading/Writing
- Replicas
- Temporary tables
- Querying

Data Bases

SQL

- Simple (SELECT FROM WHERE) queries
- HAVING, GROUPBY closures
- Joins
- Window functions

Administration

Administration

Demo services development

- flask or/and django (python)
- Virtual machines
- Containers (Docker)

Administration

DevOps (MLOps in our case!)

- Reproducible experiments (DVC)
- Service development (Mlflow, Kubeflow, etc)
- Model Versioning
- Model Monitoring (Greate expectations, SageMaker, etc)

ETL, Pipelines

ETL, Pipelines

Python pipelines

- Scikit-learn pipelines
- Construct a pipeline from the given estimators (name, transform)
- Construct a feature union from the given transformers

ETL, Pipelines

Airflow

- Schedule and monitor workflow
- More info: https://airflow.apache.org/

ETL, Pipelines

MLflow

- An open source platform to manage the ML lifecycle, including experimentation, reproducibility, deployment, and a central model registry
- Model tracking
- Model deploying
- Model registry
- More info: https://mlflow.org/

ETL, Pipelines

Kubeflow

- Deployment of machine learning (ML) workflows on Kubernetes Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications
- More info: https://www.kubeflow.org/ https://kubernetes.io/

Visualization, Dashboarding

Visualization

Python

- Matplotlib
- Seaborn
- Plotly and Dash

Visualization

BI visualization tools

- Tableau, Looker, etc

Distributed Computing

Distributed computing

Distributed filesystems

- HDFS
- Data storing, partitioning

Computing

- MapReduce
- Spark, SparkML
- HQL (Hive)
- Pig
- …

Cloud Platforms

Cloud platforms

Remote work

- Code and data transferring
- Jupyter hub
- Keys generation
- ssh, scp, …
- Session managers: tmux, screen

Cloud platforms

Amazon, GCP, Digital Ocean

- Virtual servers
- Dedicated servers
- Data storages (S3, etc)
- ML platforms and tools (AzureML, Kubernetes, Sagemaker)

Infrastructure and tools

To take away

1. Operation systems
2. **Code repository**
3. **Software engineering**
4. Administration
5. Data bases, SQL
6. ETL, **Pipelines**
7. **Visualization, dashboarding**
8. Distributed computing
9. Cloud Platforms

Машинное обучение: базовые концепции машинного обучения

Спасибо!

Эмели Драль