Data 301 Wrapup

DATA 301

Generic workflow

1. What do you want to do? predict? Cluster? Something else?

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations

Generic workflow

1. What do you want to do? predict? Cluster? Something else?
2. Get data proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations

Generic workflow

1. What do you want to do? predict? Cluster? Something else?
2. Get data proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?
3. Preprocess the Data you will spend a lot of time here, clean NaN's, normalize fields, ensure strings are consistent (VA,Va, Virginia etc..), encode ordinal data, etc…

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations

Generic workflow

1. What do you want to do? predict? Cluster? Something else?
2. Get data proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?
3. Preprocess the Data you will spend a lot of time here, clean NaN's, normalize fields, ensure strings are consistent (VA,Va, Virginia etc..), encode ordinal data, etc…
4. Explore the Data plot it, are there anomalies? can you see patterns? How to plot if you have >3 features?

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations

Generic workflow

1. What do you want to do? predict? Cluster? Something else?
2. Get data proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?
3. Preprocess the Data you will spend a lot of time here, clean NaN's, normalize fields, ensure strings are consistent (VA,Va, Virginia etc..), encode ordinal data, etc…
4. Explore the Data plot it, are there anomalies? can you see patterns? How to plot if you have >3 features?
5. Model Data build, fit and validate a model

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations

Generic workflow

1. What do you want to do? predict? Cluster? Something else?
2. Get data proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?
3. Preprocess the Data you will spend a lot of time here, clean NaN's, normalize fields, ensure strings are consistent (VA,Va, Virginia etc..), encode ordinal data, etc…
4. Explore the Data plot it, are there anomalies? can you see patterns? How to plot if you have >3 features?
5. Model Data build, fit and validate a model
6. Evaluate Model not good enough? Go to step 2, consider ensembling multiple models, consider simple default model to compare against

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations

Critical Bits

1. Domain Expertise is essential. For instance:
   - When evaluating medical images for tumors, it helps if you can recognize a tumor.
   - When exploring clustered botanical data its useful to be able to verify that groupings make sense.
   - When checking engine sensor data, it helps if you have some idea of what typical running parameters are and how anomalous behavior presents.
2. Communication and presentation skills are the most important of all.
   - If you cannot convince stakeholders to follow you, then you have wasted your time
   - Take a public speaking or acting class

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations

What we covered

General project workflow
Data preprocessing, cleaning, EDA, data leakage, Pandas
Clustering
Splitting a dataset
Handling Dataset imbalance
Regressors verses classifiers
Linear regression
Decision Trees
Random Forest
Gradiant Boosted Trees
What tree based algorithms cannot do that regressions and Neural Networks can
Explainability
Hyperparameter tuning
Cross validation
Algorithm evaluation metrics(accuracy, Precision, Recall F1, R squared)
Time series analysis

See course website for complete list

See https://neptune.ai/blog/best-practices-for-data-science-project-workflows-and-file-organizations

Where to go after this course

- SQL – there is a lot of data in databases
  - Take a Database class- at least an online one
- Do you need to scale your compute?
  - Yes. You don't use a laptop. At a minimum a local server with a GPU (or GPUs)
  - What if your data is huge and will not fit in memory? Next step is something cloud based like pyspark, Dask or terality (this area is changing fast). Take a class, or a tutorial offered by these companies.
- You have to learn to use Linux
- Start creating your own projects
- Participate in Data Science competitions (Kaggle etc.)
- Learn to use generative AI

Where to go after this course

- Its important to know how to set up cross validation and hyperparmeter tuning
- Its important to automate processes through pipelines and packages
- Learn some Neural network architectures
  - Image data- conv nets (many, many flavors)
  - Sequential data (like text processing) – use transformers (better than RNN’s or LSTM’s)
- Don’t count out linear and logistic regression- fast, explainable, understood by many

The End