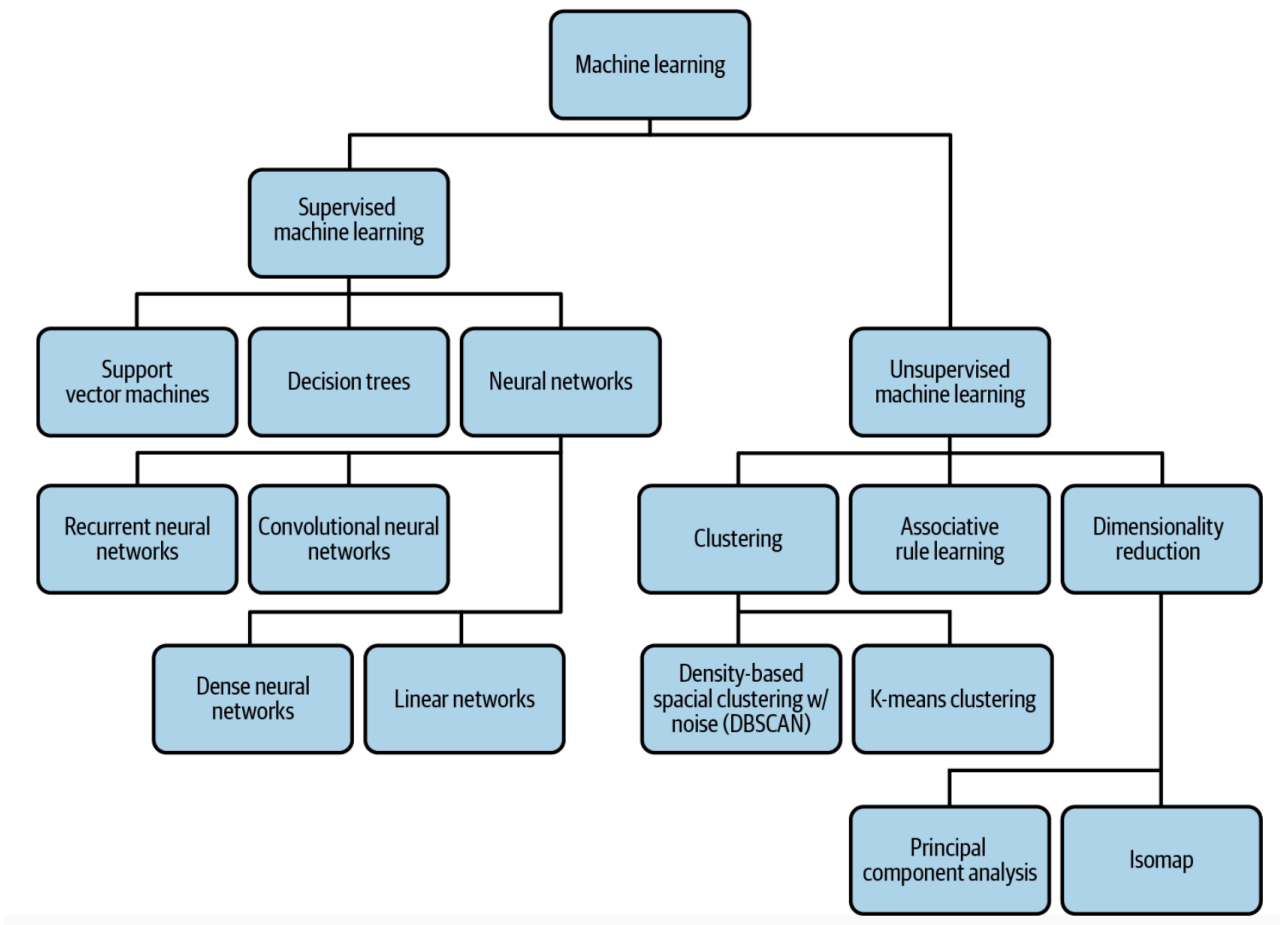
Chapter 1 The Need for Machine Learning Design Patterns

1/25/21 9:20 PM

Machine Learning Terminology

Machine Learning Models

- Algorithms that learn patterns from the data



Data and Feature Engineering

- Structured Data: numerical and categorical data
  - Alternative term: tabular data
- Unstructured Data: not represented in clean forms
  - Text, images, videos, audio
- Data validation, a process of
  - Computing statistics of the data
  - Understanding the schema
  - Evaluating dataset to identify problems

Machine Learning Process

- Training -> Evaluation
  - Serving: accepting incoming requests and sending back predictions by deploying the model as a micro-service
- Prediction: send new data to model and make use of its output
  - Online prediction: Get predictions on a few examples in a near real time
  - Batch prediction: generate predictions on a large set of data offline
- Streaming: new data ingested continuously and need to be processed immediately
  - Need multi-step solutions for feature engineering, training, evaluation and prediction
  - This is called ML pipelines

Common Challenges in Machine Learning

Data Quality

- Data accuracy
  - Need to know where data comes from and potential error in collection process
  - Need through screening for mistakes or duplications
- Data completeness
  - Model fails if new data falls out of training field (dog pictures on models trained only on cat)
  - Make sure training set has a varied representation on each of labels
- Data consistency
  - Inconsistent features: e.g. change of measuring criteria in one variable
  - Inconsistent label: e.g. labeler bias for sentiment classification
  - In data format: e.g. different spellings of same location address
- Tinelineness
  - Latency between event occurrence and when added to database
  - Might record timestamps and account for the gaps when performing feature engineering

Reproducibility

- Model given same training set might have slightly different results (randomness in training)
- Useful to set a random seed at the beginning
  - `tf.random.set_seed(value) (Tensorflow)`
  - `from sklearn.utils import shuffle`
    `data = shuffle(data, random_state=value) (Sklearn)`
- Machine Learning framework dependencies
- Model's training environment

Data Drift

Data drift refers to the challenge of ensuring your machine learning models stay relevant, and that model predictions are an accurate reflection of the environment in which they're being used

- Perform EDA to catch data drift
- Continuously update training set and retrain

Scale

- Model prototype is entirely different from infrastructure necessary to a production model for large size of data
- Machine Learning Engineers are in charge of determining necessary infrastructure for a specific training job (e.g. GPU, image models, etc.)