Notes on: "Hands-On Machine Learning with Scikit-Learn & TensorFlow"

Alexandre De Spiegeleer

October 19, 2020

Chapter 1

The machine learning landscape

1.1 Categories of ML

Different categories of ML algorithms:

Properties	**Descrition**	**Examples**
Supervised	Need laballed training data	kNN, Linear Reg., Logistic Reg., SVM, Decision Tree, ...
Unsupervised	No labelled data, the algorithm puts similar data together	Clustering: k-means, HCA. Dimensional reduction: t-SNE
Semi-Supervised	Both labelled and not labelled data	Deepd belief network, restricted Boltzman machine
Reinforcment Learning	In the learning process give rewards or penalites depending on the action done	
Batch Learning	Trains using ALL data → If needs to retrain, need to train on all the data again	
Online Learning	Train by mini-batches: Train incrementally and can start again from a previous minibatch run	
Instance Based	On new data, check distance to known data and assign same output	
Model-based	Use the data to make predictions / interpolates to new data	

1.2 Challenges

1.2.1 data

Not enough data	Simple problems → min. thousands of examples.
Non representative data	Data used for training must include similar data to those for predictions. (Poor extrapolation capacity)
Sampling bias	if the sampled data are not representative.
Poor quality data	if errors, outliers and noise in data
Irrelevant features	There should be enough relevant features and not much crap.

1.2.2 algorithm

Overfitting	→ Lose predictability. Occurs because model too complex compared to data.
Underfitting	Model too simple compared to the data to be represented

- *Regularization*
  → Making a model simpler to avoid
- *Hyperparameter*
  → Parameters of the learning algorithm that dictates the amount of regularization.

1.3 Evaluating performance

Split the data into training set (80%) and testing set (20%). Evaluation on the test set gives an estimate of the error on unseen data. When training multiple models on the training set and testing them on the test sets, our selection of the "best" model is which model best fits the test data set. → Model cannot necessarily be good on new data. Thus, keep a validation set.

1. Train multiple models with different Hyperparameter on the training set
2. Select the model that performs best on the test set.
3. Verify that the model is indeed good on the validation set

It is common to split the training set to train several models on sub-sets of the training set and train one final model on the whole training set.

Chapter 2

End-to-end machine learning project