Hands-on Machine LEarning with Scikit-learn, Keras & Tensorflow

Aurelien Geron

Second EDITION

OREILLY

The Machine Learning Landscape

Chapter 1 solving Exercises of Hands-On ml with scikit-learn, Keras, and TensorFlow Book Edition 2.

Abdelrahman Rezk

Q-1

How would you define a machine Learning ??

Is the way that computer trying to learn from Experiment instead of writing each instructures for each new rule, and by the way we can not assume all of the test cases, also machine learning around us in different ways that we do not notice, its not just about *robuts*, no its a general meaning of other application like spam classification, price prediction, OCR, recommendation system, voice recognition, and others.

Also, Machine learning is the field of study that gives computers the ability to learn without being explicitly programed.

Also, A computer program is said to learn from experience **E**, with respect to some task **T**, and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**.

- E ==> When model train on training dataset like spam and ham mails.
- T ==> The prediction of an e-mail as spam or ham.
- P ==> How well model predict spam verse ham.

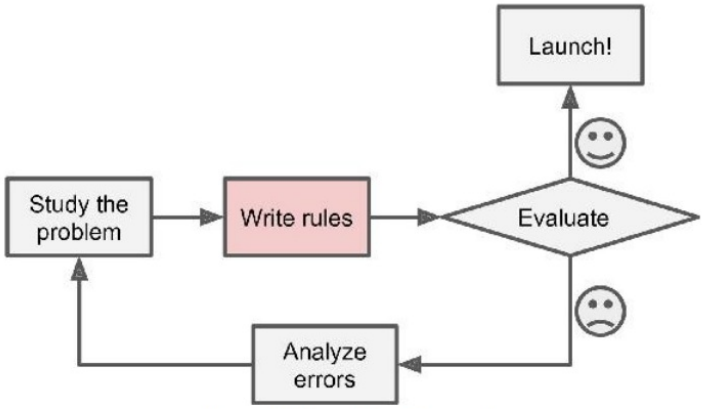

Figure 1-1. The traditional approach

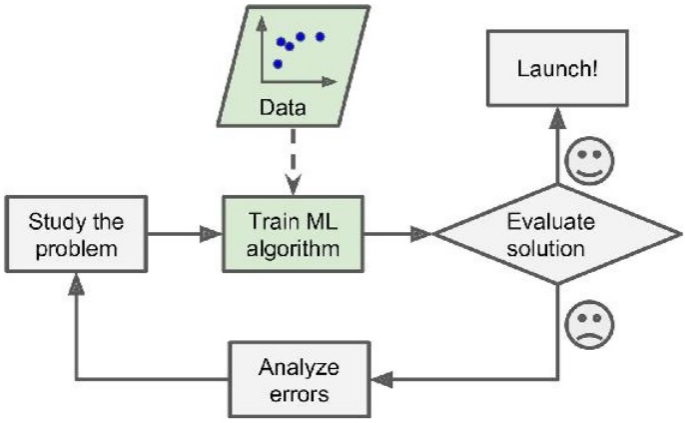

Figure 1-2. Machine Learning approach

Q-2

Can you name four types of problems where it shines ??

Some of the problems that machine learning shines in:

- Voice Recognition [word lile **two**]
- Summarize text
- Recommendation System
- Classification

Q-3

What is a labeled training set ??

The machine learning models have different types one of these types is **Supervised Learning** where models have some of the training examples to learn from and we till the model the answer for each of them in the training set, we called these answers are labeled training set.

Q-4

What are the two most common supervised tasks??

As we talked above the ml have different types one of them is **Supervised** and the supervised its self contain two main types or solve two popular problem which are **Linear Regression** and **Classification or Logistic Regression**

- Linear Regression
- Logistic Regression

Q-5

Can you name four common unsupevised tasks??

The ml models as contain **supervised Learning** it also contain type which is **Unsupervised Learning**, it's about let the model learn without any labeled data, and the unsupervised learning solve different tasks:

- Anomaly Detection & novelty detection
- Visualization & Dimensionality Reduction
- Clustering
- Association rule learning

Q-6

What type of machine learning algorithm would you use to allow a robot to walk in various unknown terrains??

- Reinforcement learning

Q-7

What type of algorithm would you use to segment your customers into multiple groups??

- Unsupervised Learning Approach using **Clustering Algorithm**

Q-8

Would you frame the problem of spam detection is supervised learning problem or an unsupervised learning problem??

The problem is the **Supervised Learning** problem because it depends on other training sets you have labeled before of flag another task.

Q-9

What is the online learning System??

Before we start with an online learning system we need to know which the other approach of a system which is the **Batch Learning** in the system work like that, the model start to learn from all available data then take the weights and put the model to production and like this models, should be trained from scratch if your system starts to deal with another type of data, and like this process is consuming your CPU, i/o, network, and other resource and actually a lot of time with each new types of data you need to run the model from scratch on old and new data.

Instead of the previous case which is **offline Learning** we can run the model to work incrementally with the new data, by feeding the data sequentially, either one instance for each time or group of data which is called mini-batch, this type of work is so good to work with because the system is still updating your parameters online as a new type of data coming and here should we take care of our **learning rate** to be suitable not increase the weights in high or decreasing in low, should we make the trade-off here because it will affect our system in production mode.

Q-10

What is the out-of-core learning System??

It's about the Batch Learning but in the case that we can not deal with a huge amount of data which can not be feeding to our memory, or the resource we have, so in like this case online learning can dealing with, if we have a simple resource because its run mini-batch, and here is the model runs training step on the data and repeats the process until it has run all of the data.

Q-11

What is the type of algorithm relies on similarity measure to make prediction. ??

This type is:

- Instance-Base Learning

This work as we can imagine that, some of the spam e-mail contains some of the words common in, and the new e-mail will be generalised as spam if the similarity of the new email with olds spams one are common of the words in. and the other architecture is what we hear the model-based learning.

Q-12

What is the difference between model parameters and learning algorithm hyperparameters??

We can simplify this by that, model parameters that something the model control of and by learning on the data the model change these values to the values that fit very well on the data, other ones which are the hyperparameters are those you need to set before the model run on the training data and you should fine-tune these hyperparameters very well.

Q-13

What do model-based learning search for? what is the most common strategy they use to succeed? how they make prediction??

The model-based search for the pattern in the data by updates the weights of the model for each time in the learning on the training data, and it follows the strategy that, these weights that updates by the model should decrease or minimize the cost function for the future prediction when model be on the production, at the end we save these weights which minimize the cost function to the global minium on the new data.

Q-14

Can you name four of the main challenge in machine learning??

The two things that can go wrong are **bad algorithm** and **bad data** for bad data:

- Insufficient quantity of the training data.
  - Model not like child can learn by some of the example the dads talking about, its need thousands or millions of examples.
- Nonpresentive training data
  - Data do not representative new types of data or not all samples of the whole data.
- Poor Data.
  - When your data is full of outliers, errors, and noise.
- Irrelevant attributes or features.
  - A model with irrelevant features can lead the model to bad predictions in the future.

bad algorithm is this leading to overfitting and underfitting problems.

Q-15

if your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions???

This problem is called Overfitting problem, which the model work well on the training data and fit the training data very well but can not genrlize the learning so in the testing data the error being large, we can handle like these problem:

- Select a model with fewer parameters.
- Increase the size of the training dataset.
- Fix data errors and remove outliers.
- Make a regularization part for the model to reduce the risk of overfitting.
- Ignore some of the features or features selection.

Q-16

What is a test set and why would you want to use it ???

The test set is the part we evaluate the model on before we launch to production, we can not know how our model is by the words, we should evaluate the model on the test set which have not seen before to ensure that the model can work on the production.

Q-17

What is the purpose of a validation set ???

A validation set is used to compare models. It makes it possible to select the best model and tune the hyperparameters.

Its when the test set is not enough in the case of having a different model that works well on the problem, the validation set helps me decide which of the model should use.

The holdout validation is the way we hold out part of the training set to evaluate several models and select the best one, the new held-out set is the validation set or called development set or dev set, then train you best model on, then training the model on the full training set after tuning your hyperparameters on dev set, then you also should test your model by the test set.

Q-18

What can go wrong if you tune hyperparameters using the test set???

We can not knowing how well our model will work in the real-life or production.

The train-dev set is used when there is a risk of mismatch between the training data and the data used in the validation and test dataset(Which should always be as close as possible to the data used once the model is in production).

The train-dev set is a part of the training set that's held out (the model is not trained on it).

The model is trained on the rest of the training set and evaluated on both the train-dev set and the test(called also validation set).

If the model performs well on the training set, but not on the train-dev set, then the model is likely to overfit the training set.

If the model performs well on both the training set and train-dev set, but not on the test(validation) set, then there is probably a significant data mismatch between the training data and the validation + test data, and you should improve the training data to make it look more like the validation + test data.

Q-19

What is cross-validation and why would you prefer it to a validation set ???

Is a way that the model test on all the dataset by divide the data into many small validation sets, each model is evaluated once per validation set after it is trained on the rest of the data.

If you tune hyperparameters using the test set, you risk overfitting the test set, and the generalization error you measure will be optimistic. So you may launch a model that performs worse than expected.