Questions Chapter 1

Loris Widmer

Definition and Motivation of Machine Learning

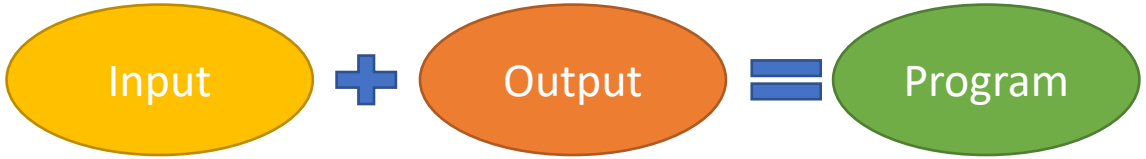
- Task
  - The task is to estimate the weight of a person based on his/her height.
- Performance measure
  - The performance measure is used to examine how accurate the estimation is. The ratio of correctly estimated weight can be choosen to evaluate the accuracy.
- Training set
  - The training set in this example would be a number of pairs of Hight and weight(the desired solution) which are fed to the algorithm. Based on this training set the system learns what weight is expectet with a certain hight.
- Training instance
  - Each training example is a training instance. In our case each pair of hight and weight: 180 cm – 80 kg.
- Model
  - The model is the part of the machine learning system that learns and examines the weight based on the hight.

Difference between traditional programming and machine learning

Traditional Programming

- Manual process
- Program behaves acording to programmed rules
- Potentially very long program
- Complicate maintenance
- Changing environment needs manual re-programming

Machine Learning



- Programm learns based on training
- Much shorter program
- «Easy» maintenance
- More accurate
- Recognizes changing environment

- Motivation for machine learning
  - Better fit for complex problems
  - Adoption to changing environment
  - Recognises patterns
- Examples for approaches
  - Traditional programming
    - If a task is very clear: calculating the birthyear based on the age of a person
  - Machine Learning
    - Where the environment changes: Suggesting pictures on Instagram(preferences of a person might change)
- Strenghts of machine learning
  - If a system runs for long enough and has enough input, it digs into the data (data mining) and checks if there is a new pattern to consider. This allows the system as well to react to fluctuating environments and adapt to new situations.

Types of Machine Learning Systems

- Supervised learning
  - The training set provided to the algorithm has the input and the wanted output (labels)
- Unsupervised learning
  - The input of the training set does not have any labels included. The algorithm has to learn the answers without the help of the programmer
- Self-supervised learning
  - Initially an unlabeled dataset is used to generate a fully labeled dataset.
- Semi-supervised learning
  - Only a part of the training set is labeled.
- Reinforcement learning
  - The learning system will get rewards or punished based on the answer it gives. A policy defined what the correct actions are. The system then tries to get as much reward as possible

- Difference between batch learning and online learning

Batch learning (disadvantages red)

- System incapable of incremental learning
- Must use all data to train
- Time and computing intense
- Done offline
- Once launched, it runs without training

Online learning

- System is trained incrementally
- Data is feeded sequentially
  - Individual or in small groups
- Each step fast and cheap

- Models which do batch learning
  - Linear regression

- Out-of-core learning
  - Out-of-core learning is used during online learning. If the dataset is too big to fit in one machine's main memory, it is done incrementally. It repeats the training step with parts of the dataset. (not done on the live system)

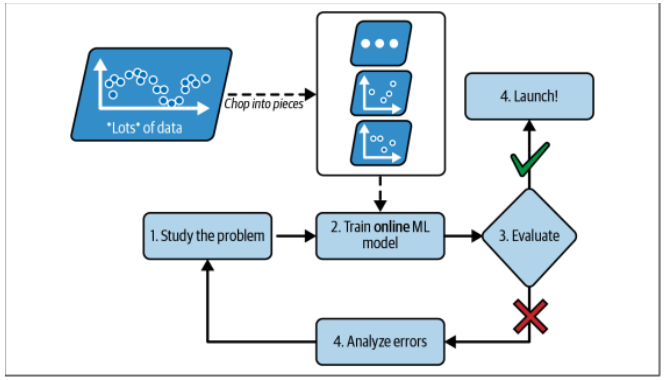


Figure 1-15. Using online learning to handle huge datasets

- Challenges of online learning
  - If bad data is fed into the system, the performance will decline
  - Either from a bug or a "hacker"
  - System needs to be monitored and in case of performance drop learning has to be turned off

The main challanges in Machine Learning

- Challenges with Machine Learning
  - Bad training data quality
  - Nonrepresentative training data
  - Not enough training data
  - Irrelevant Features
  - Overfitting/Underfitting
  - Stepping back

- Sampling noise/Bias
  - If the sample is too small, there will be sampling noise. Nonrepresentative data by chance. If there is a very large sample but the method is flawed, there will be sampling bias. This is as well nonrepresentative.

- Features
  - Features are used by the system to make predictions. If a system must predict a value based on the input X, Y and Z, the letters X, Y and Z are features.
- Feature selection/feature extraction
  - Feature selections is selecting the most useful features, while feature extraction the combination of existing features to a better one is.
    - During Linear regression, the least important features (least impact on output) can be excluded with the help of a ranking.
    - Feature extraction can be achieved with the help of patterns in the data. Features can be combined to a more informative one.

	Overfitting	**Underfitting**
What is…	Overgeneralizing, works well on training data, does not generalize well.	The model can't learn the underlying structure of the data, poor performance on training and test data
What favours…	Small of noisy training sets, to complex model	Model is too simple
How to prevent…	Simplfy model (regularization), more training data, reduce the noise in training data	More powerful model/more parameters, better features, reduce constraints, more training data

Testing and Validating

- Motivatin of holdout validation
  - Holdout validation is a solution to the problem of training a model to perform best on a particular training set.

- Test set
  - Sample of data which is used to provide an unbiased evaluation of final model
- Training set
  - Full training set minus the validation set, used to fit the model
- Validation set
  - Held out set of the training set

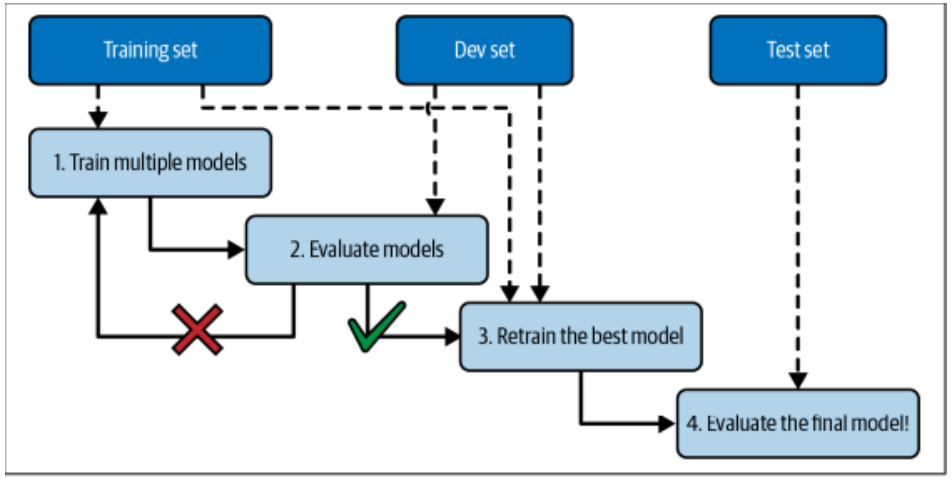


Figure 1-25. Model selection using holdout validation