SE Digital Organizations: Assignment 1

Tobias Höpfl

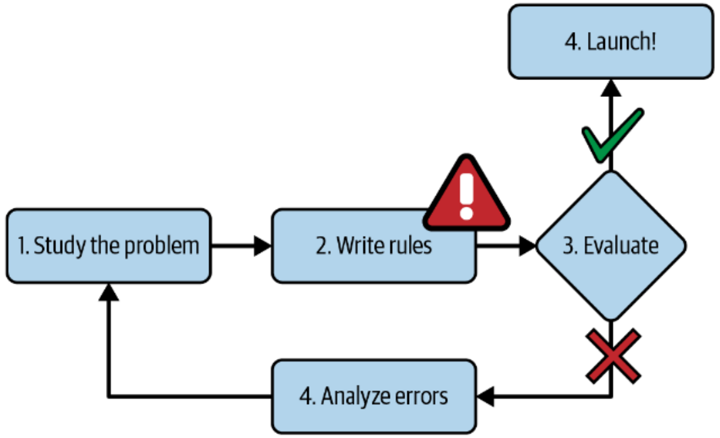
Definition and Motivation of Machine Learning

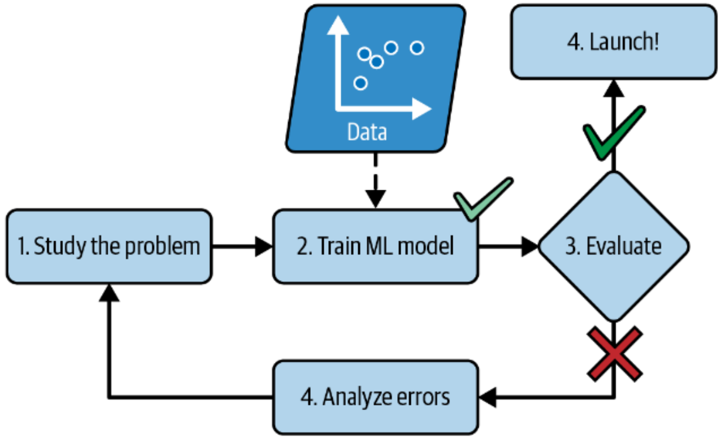
Read through the definition of machine learning. You want to estimate the weight of a person based on his or her height. Now try to describe the terms task, performance measure, training set, training instance, and model using this concrete example.

- Task: estimate the weight of a person
- Performance measure: e.g., mean deviation from the actual height
- Training set: collection of weights and heights (as labels) of people that is used to train the model
- Training instance: weight and height of an individual used to train the model
- Model: Makes the prediction, e.g., linear regression

What is the difference between the traditional programming and the machine learning approach? What motivates the machine learning approach? When will I use one approach and when will I use the other, think about an example. What other strengths does a machine learning approach have? Try using the terms "fluctuating environments" and "data mining".

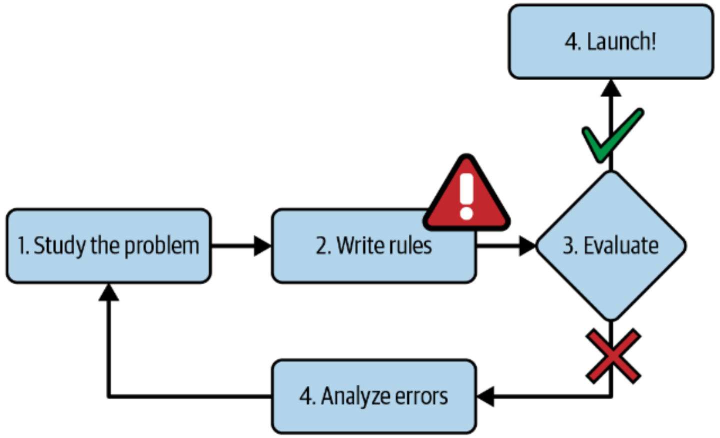
- The traditional programming approach is based on rules, which require a lot of fine-tuning in the code (time-consuming)
- Problems: Some tasks are too complex. In fluctuating environments data changes rapidly. In traditional programming all the hard coded rules would have to be reprogrammed
- Machine learning: algorithm learns by itself and can match the new environment (see especially online learning ("on-the-fly"))

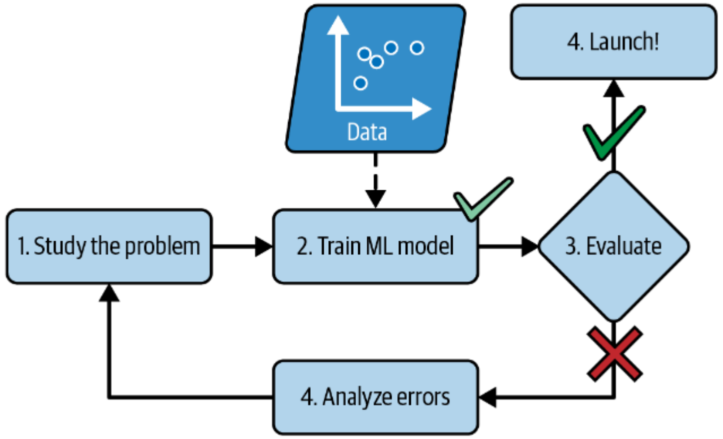




- Other strengths:
  - ➢ Tasks where no algorithm exists (because too complex)
  - ➢ Can help humans learn from data by looking at the trained ML model (Data mining)

- Examples:
  - ➢ Classifying documents by using a few well-defined business rules based on the date, author and category → Traditional programming approach (simple problem)
  - ➢ Classifying pictures of furniture→ ML approach (complex problem, style can change with time)





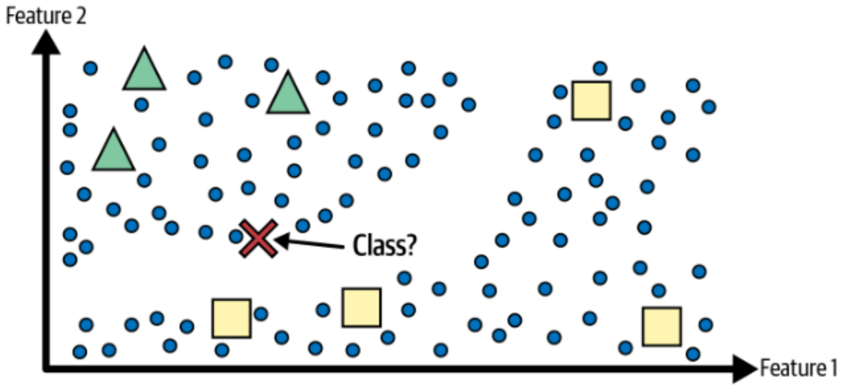
Types of Machine Learning Systems

Try to summarize all five training supervisions (supervised learning, unsupervised learning, self-supervised learning, semi-supervised learning, and reinforcement learning).

- Supervised learning: Training set includes the desired solution (label)
- Unsupervised learning: Desired solution not included, algorithm learns on its own (e.g., clustering)
- Self-supervised learning: Generate fully labeled dataset from fully unlabeled (e.g., image completion, mostly useful for another task)
- Semi-supervised learning:

  Some labeled and some unlabeled data (see graph)
- Reinforcement learning: Learn from the environment

  in the form of rewards and punishments (e.g., game bot)



What is the difference between batch and online learning? What is the disadvantage of batch learning? Do you already know of models that definitely do batch learning? What does the term “out-of-core learning” mean? What are challenges in online learning?

- Batch learning:
  - train the model once in the beginning
  - disadvantage: as data changes, the model degrades
  - cannot learn incrementally
  - Retraining the model: time-consuming
- Online learning:
  - can learn incrementally (“on-the-fly”) from new mini-batches of data
  - Challenges: Risk to retrain model online, if it is bad data (live system will have deteriorating quality) → models in highly security critical areas will use batch learning
- Out-of-core: If a data set is too big and cannot fit in the main memory (feed the model iteratively only parts of the data)

The main challenges in Machine Learning

Briefly name all the challenges we encounter in Machine Learning

- Insufficient Quantity of Training Data
- Nonrepresentative Training Data
- Poor-Quality Data
- Irrelevant Features
- Overfitting the Training Data
- Underfitting the Training Data

The main challenges in Machine Learning

The data are the basis for our models. If the data is not representative, we will only be able to make limited predictions with our model. What do we mean by sampling noise and sampling bias?

- Sampling noise: small samples, nonrepresentative data by chance
- Sampling bias: data in large samples can be systematically non-representativene

  (because the sampling method can be flawed)

What are features? Use a simple ML model and describe what is meant by feature selection and feature extraction?

- Features: characteristic property in the data (e.g., weight, height and cloth size of children)
- Feature selection: select most useful features for training the model for a specific task (e.g., for estimating the age of the child, the height might be the most relevant feature and the others can be ignored)
- Feature extraction (cf. dimensionality reduction): simplify data without losing information (e.g., cloth size might have a very high correlation with weight and height, so it can be left out without losing much information)

What is and what favors overfitting in Machine Learning and how can we prevent it? What is and what favors underfitting in Machine Learning and how can we prevent it?

- Overfitting: Performs good on the training data but off on real data (see graph), maybe too many parameters
- Underfitting: Too simple to capture the underlying structure of the data
- Prevention: Test data, to evaluate performance
- Less degrees of freedom for overfitting
- More powerful model needed for underfitting

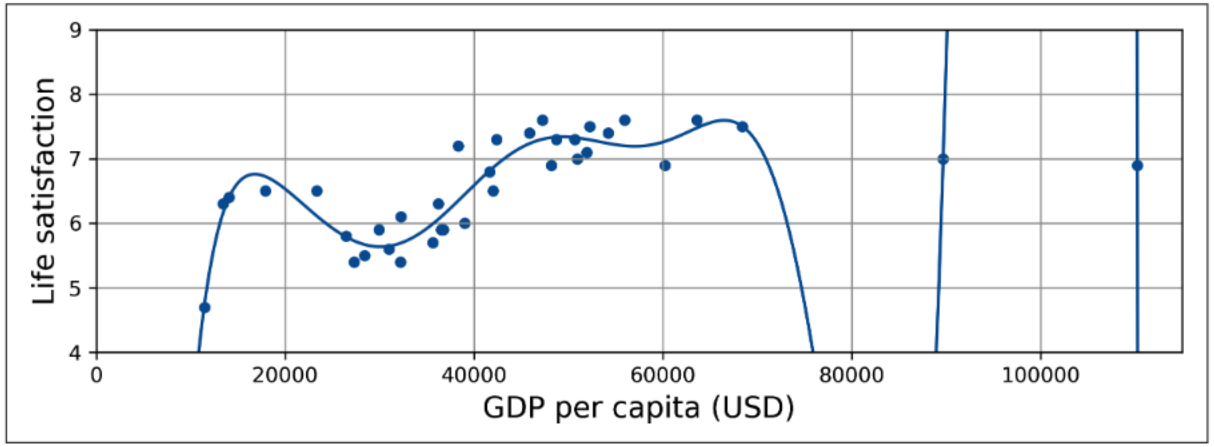


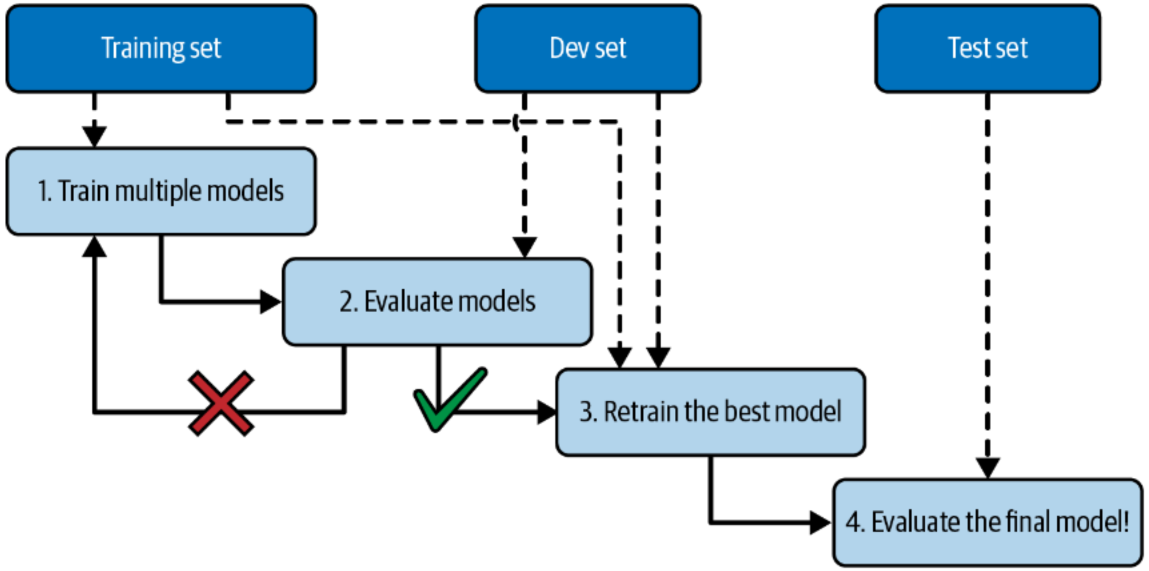
Figure 1-23. Overfitting the training data

Testing and Validating

What is the motivation of holdout validation? Using Figure 1-25, try to explain what is the test set, the training set, and the validation set?

*Motivation: find the right model (*e.g., when different hyperparameter are under consideration)

- Training set: Used to train multiple models
- Dev set (=validation set): Used to find the best model
- Test set: Evaluate final model after it was trained again on the combination of training and dev set


Training set
Dev set
Test set
1. Train multiple models
2. Evaluate models
3. Retrain the best model
4. Evaluate the final model!