The Machine Learning Landscape

Questions Chapter 1 - Benedikt Kneussl

Definition and Motivation of Machine Learning

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.

In this task, the objective is for a machine to estimate weight based on height, following its learning process. To evaluate the performance of the machine learning algorithm, defined performance measures are used. For instance, an acceptable weight range for a given height can be established. This means that an estimated weight of 110 lbs should not correspond to a height of 6'4", while it could be acceptable for a height of 5’2" to 5’4".

To train the system, a set of height/weight combinations would be provided as training sets. Each individual height and weight combination within the training set would then become a training instance. For example, “6’1”, 180 pounds” would be a training instance for a male person in our example.

Models, such as neural networks or random forests, are programs that learn, identify patterns, or create predictions based on datasets used during the training process. The machine learning model is the outcome of this training process. Ultimately, the machine learning algorithm will be able to estimate weight based on height, after the training process has been completed.

Definition and Motivation of Machine Learning

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.

> 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*.
>
> —Tom Mitchell, 1997

Your spam filter is a machine learning program that, given examples of spam emails (flagged by users) and examples of regular emails (nonspam, also called "ham"), can learn to flag spam. The examples that the system uses to learn are called the *training set*. Each training example is called a *training instance* (or *sample*). The part of a machine learning system that learns and makes predictions is called a *model*. Neural networks and random forests are examples of models.

In this case, the task *T* is to flag spam for new emails, the experience *E* is the *training data*, and the performance measure *P* needs to be defined; for example, you can use the ratio of correctly classified emails. This particular performance measure is called *accuracy*, and it is often used in classification tasks.

Machine Learning vs Traditional Programming

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 **machine learning** approach differs from **traditional programming** in its ability to automatically learn patterns and adapt to changing inputs, making it a more maintainable and precise solution for complex problems.

Traditional programming often requires the manual creation of complex rules and struggles to detect typical phrases or patterns in large datasets. raditional programming is better suited for projects that require manually created algorithms to solve a problem.

In contrast, **machine learning** utilizes data mining to learn patterns based on frequently used phrases, making it better suited for projects that require quicker results and fewer adaptations. This approach is commonly used in applications such as speech recognition and reaction in programs like Siri or Alexa.

In summary, the machine learning approach offers greater adaptability and precision, while traditional programming may be more appropriate for simpler projects. Thus, the machine learning approach's ability to handle *fluctuating environments* and use *data mining* techniques make it a powerful solution for challenging problems.

Machine Learning vs Traditional Programming

Screenshots

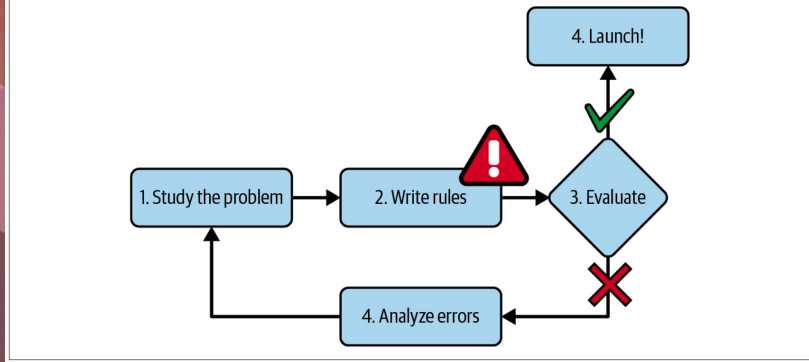

Figure 1-1. The traditional approach

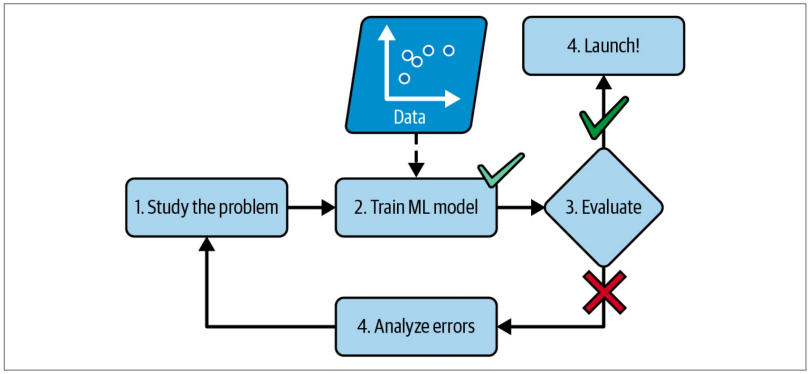

Figure 1-2. The machine learning approach

To summarize, machine learning is great for:

- Problems for which existing solutions require a lot of fine-tuning or long lists of rules (a machine learning model can often simplify code and perform better than the traditional approach)
- Complex problems for which using a traditional approach yields no good solution (the best machine learning techniques can perhaps find a solution)
- Fluctuating environments (a machine learning system can easily be retrained on new data, always keeping it up to date)
- Getting insights about complex problems and large amounts of data

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: Input data provided with expected outcomes (classification, prediction, etc.) based on features.

Unsupervised Learning: Analyzes complex, unlabeled data to discover patterns and trends (often using visualization algorithms).

Self-Supervised Learning: Creates labeled dataset from unlabeled data by predicting expected output (e.g. animal classification or completing an image).

Semi-Supervised Learning: Combines supervised and unsupervised learning using partially labeled data (e.g. facial recognition, clustering).

Reinforcement Learning: Agent learns from environment by taking actions and receiving rewards/punishments to maximize rewards over time.

Types of Machine Learning Systems

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, also known as offline learning, requires the system to be trained using all available data before it can make any predictions. This process is slower and must be done offline, using large amounts of memory. One of the main disadvantages of batch learning is that it cannot learn incrementally, and it requires all available data to be present before it can make predictions.

Online learning, on the other hand, can learn and adapt independently using large datasets, making it quicker and more cost-effective. This is achieved through a process called "out-of-core learning," which allows the system to work on parts of the data one at a time, without having to load the entire dataset into memory. Online learning is particularly useful when dealing with large datasets that cannot fit into a single machine's memory.

However, online learning presents some challenges as well. The quality of the data fed into the system is critical to its performance. If the data is of low quality, the system's overall performance will suffer, leading to negative outcomes. Bad data can arise from bugs or hacking and can be reduced by monitoring the system closely.

Types of Machine Learning Systems

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?

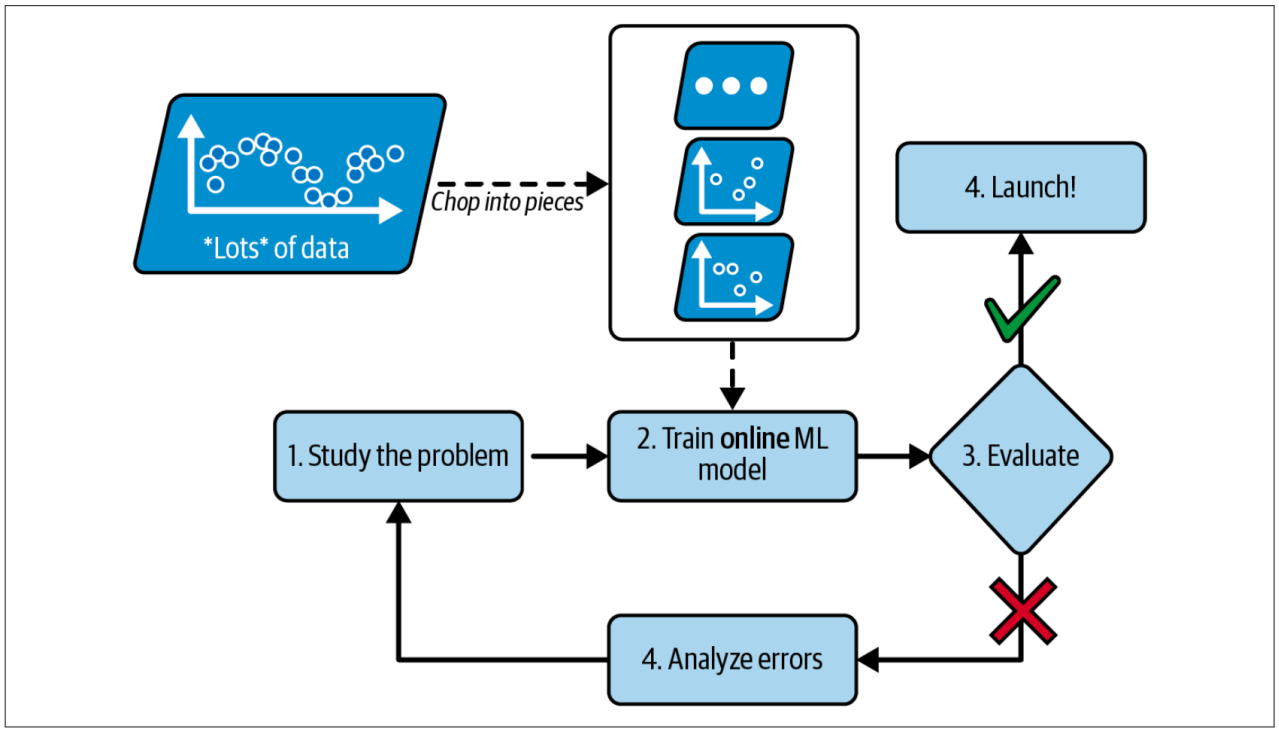

Figure 1-15. Using online learning to handle huge datasets

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
- Stepping Back

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 refers to non-representative data that results from a small sample size. This type of non-representative data occurs due to chance and can skew the results of the machine learning model.

Sampling bias, on the other hand, refers to non-representative data that results from a flawed method. This type of non-representative data can occur when the selection of data is not random, and certain types of data are overrepresented or underrepresented in the dataset. Sampling bias can occur even when working with large datasets, making it a more pervasive problem than sampling noise.

In summary, while sampling noise and sampling bias are both types of non-representative data, they have different sources. Sampling noise occurs due to a small sample size, while sampling bias results from a flawed method that leads to non-random data selection.

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?

Examples of Sampling Bias

Perhaps the most famous example of sampling bias happened during the US presidential election in 1936, which pitted Landon against Roosevelt: the *Literary Digest* conducted a very large poll, sending mail to about 10 million people. It got 2.4 million answers, and predicted with high confidence that Landon would get 57% of the votes. Instead, Roosevelt won with 62% of the votes. The flaw was in the *Literary Digest*'s sampling method:

- First, to obtain the addresses to send the polls to, the *Literary Digest* used telephone directories, lists of magazine subscribers, club membership lists, and the like. All of these lists tended to favor wealthier people, who were more likely to vote Republican (hence Landon).
- Second, less than 25% of the people who were polled answered. Again this introduced a sampling bias, by potentially ruling out people who didn't care much about politics, people who didn't like the *Literary Digest*, and other key groups. This is a special type of sampling bias called *nonresponse bias*.

Here is another example: say you want to build a system to recognize funk music videos. One way to build your training set is to search for "funk music" on YouTube and use the resulting videos. But this assumes that YouTube's search engine returns a set of videos that are representative of all the funk music videos on YouTube. In reality, the search results are likely to be biased toward popular artists (and if you live

The main challenges in Machine Learning

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

Machine learning models use features, which are independent variables, as inputs to make predictions. Relevant features are important for the quality of the system. Feature selection and feature extraction are techniques used to improve accuracy. Feature selection selects the most important features, reducing overfitting, while feature extraction transforms features into a new set of features that capture important information.
Decision trees can use information gain for feature selection, while binning can be used for feature extraction. These techniques are crucial for improving the effectiveness of machine learning models.

The main challenges in Machine Learning - Overfitting

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

Overfitting is when a machine learning model is successful in training data, but it cannot generalize outside of the training data for accurate predictions or outcomes. This issue is often linked to sampling noise, where a small sampling size and a complex model can lead to predictions based on limited data. To prevent overfitting, simplifying or regularizing the model and focusing on fewer parameters can help. Additionally, avoiding sampling noise can be achieved by increasing available data or removing data outliers.

Overfitting

Figure 1-23 shows an example of a high-degree polynomial life satisfaction model that strongly overfits the training data. Even though it performs much better on the training data than the simple linear model, would you really trust its predictions?

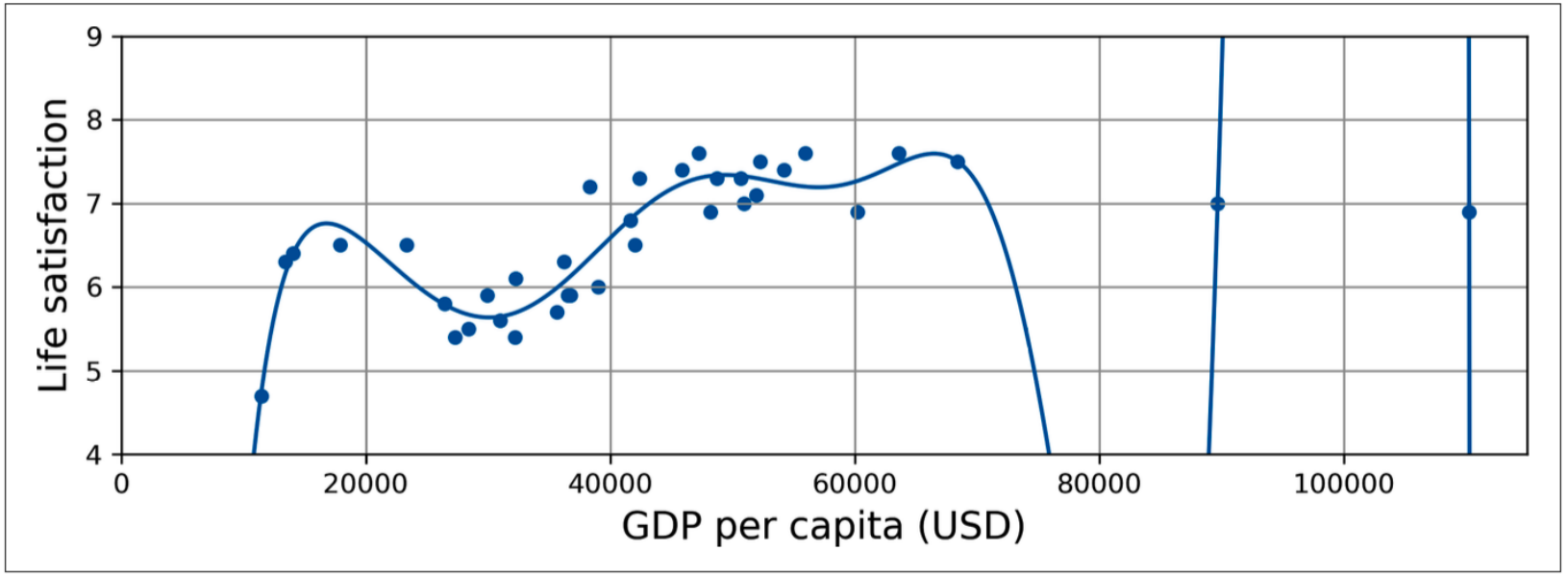

Figure 1-23. Overfitting the training data

The main challenges in Machine Learning - Underfitting

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

When a model is too simple to learn the underlying structure of the data, it is said to be underfitting. This can happen when the training is not sufficient or when the model is not complex enough to capture the complexity of the data. To prevent underfitting, one can increase the complexity of the model by adding parameters, reducing the constraints, or introducing features with improved quality.

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?

Holdout validation is the solution to finding the model that fits best to the desired outcome. By using parts of the training set on different types of models, the most appropriate model can be deducted according to the outcomes.

The training set is a subset of the data that is used to train the machine learning model. This set contains input features and corresponding output labels.

The validation set is a subset of the data that is used to evaluate the performance of the machine learning model during training. This set is used to tune the hyperparameters of the model, such as learning rate or regularization strength, to improve its performance.

The test set is a subset of data that is used to evaluate the final performance of the machine learning model after it has been trained according to the training and validation sets. This set is used to simulate the model's performance on unseen data. This set should be separate from the other two as it prevents a biased outcome and gives more accurate estimates of performance.

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?

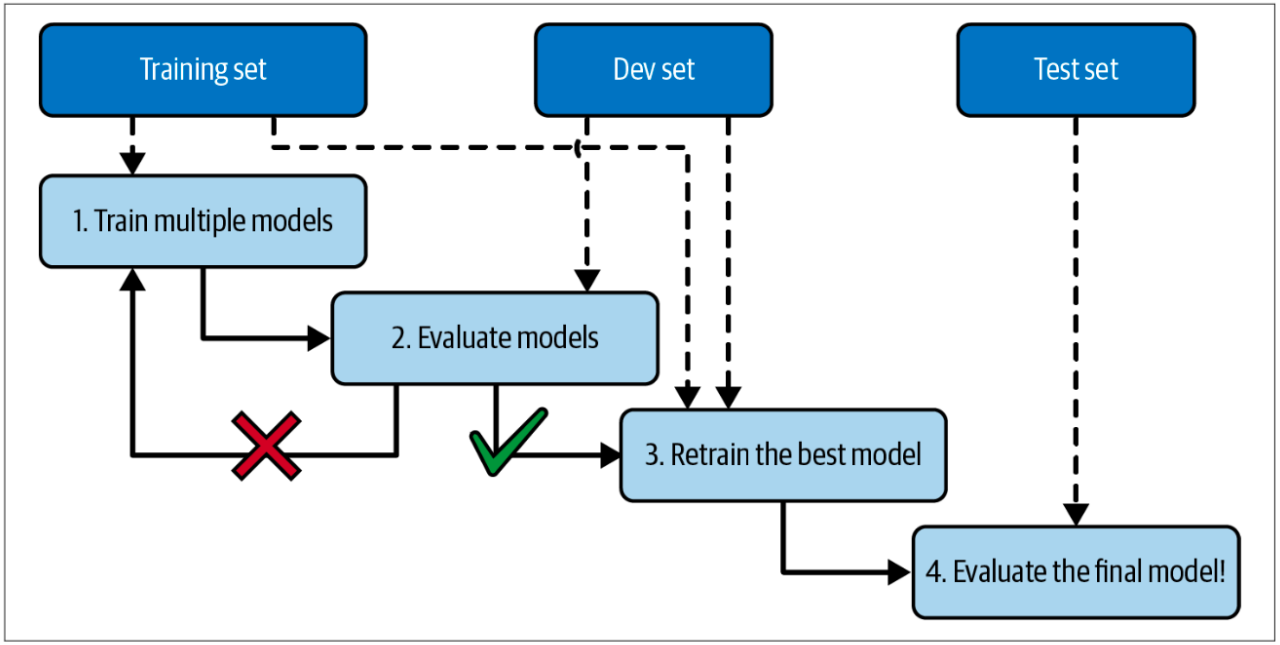

Figure 1-25. Model selection using holdout validation