Assignment 1
Title: SE Current Topics of Informations Systems, especially Digital Society

Igor Gatchin

Table of content

- **Definition and Motivation of Machine Learning**
- **Types of Machine Learning Systems**
- **The main challenges in Machine Learning**
- **Testing and Validating**

Definition and Motivation of Machine Learning

Q: try to describe the terms task, performance measure, training set, training instance, and model using this concrete example.

In this example, a **task** is to estimate the weight of a person based on his height. The performance measure could be the mean squared error (MSE) between the predicted weights and the actual weights of a group of people.

A **performance measure** in this example could be the mean squared error (MSE) between the predicted values and the actual values. The lower the MSE, the better the model's performance.

The **training set** is a subset of the dataset that is used to train the model.

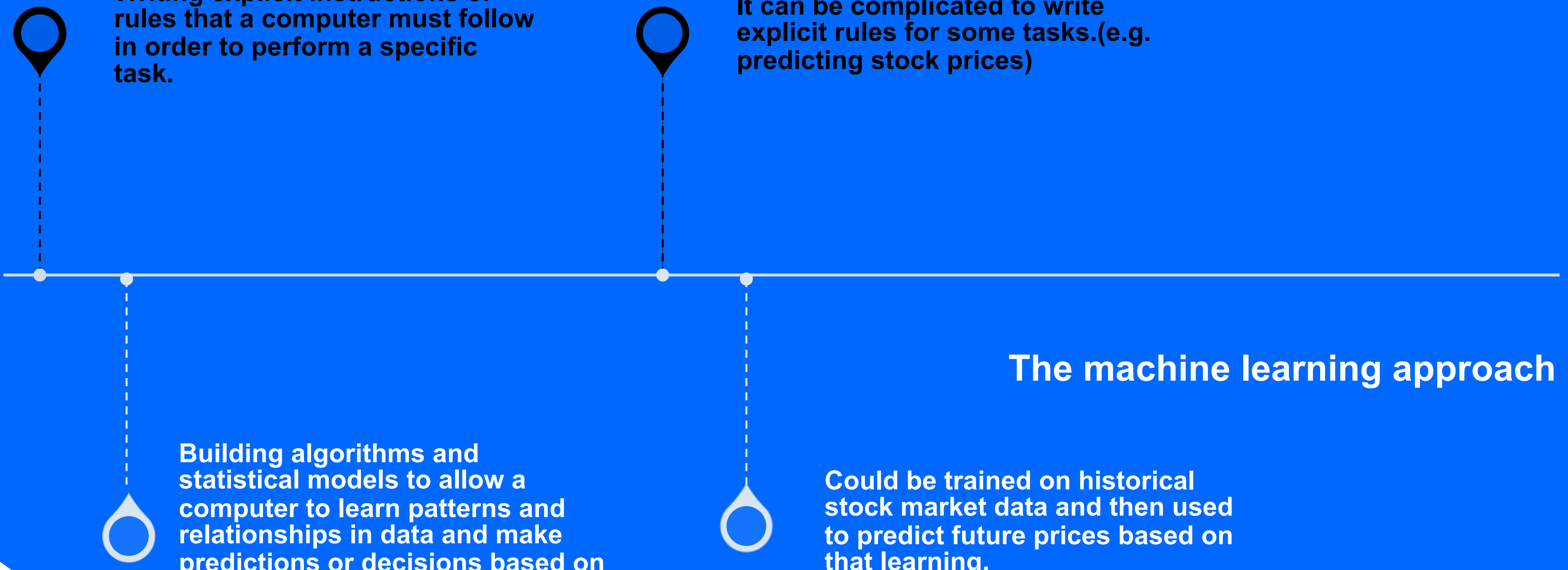
A **training instance** is a single example in the training set, consisting of a set of features and the target variable.

A **model** is the algorithm used to estimate the weight of a person based on their height. In this case, a linear regression model could be used that learns the relationship between height and weight by minimizing the MSE in the training set.

Definition and Motivation of Machine Learning

Q: What is the difference between the traditional programming and the machine learning approach? What motivates the machine learning approach?

The traditional programming approach



Writing explicit instructions or rules that a computer must follow in order to perform a specific task.

It can be complicated to write explicit rules for some tasks.(e.g. predicting stock prices)

The machine learning approach

Building algorithms and statistical models to allow a computer to learn patterns and relationships in data and make predictions or decisions based on that learning.

Could be trained on historical stock market data and then used to predict future prices based on that learning.

Definition and Motivation of Machine Learning

Q: 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

Writing explicit instructions or rules that a computer must follow in order to perform a specific task.

In the case of image recognition, traditional approach may be used to define explicit rules for identifying certain features in an image, such as edges, shapes, and colors.)

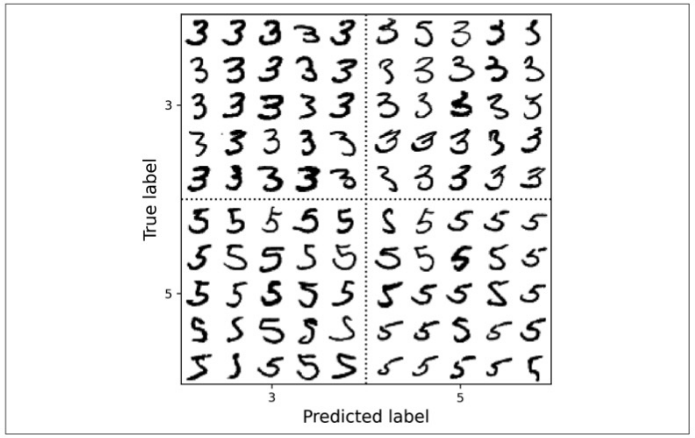


Figure 3-11. Some images of 3s and 5s organized like a confusion matrix

Strengths of using a machine leaning approach:

- ability to adapt to fluctuating environments and perform data mining

- In fluctuating environments, where the underlying patterns and relationships in the data may be constantly changing

- Machine learning is also useful in data mining, as it can automatically discover previously unknown patterns and relationships in large datasets.

The machine learning approach

Machine learning could be used to learn patterns in large datasets of images and use them to accurately classify new images

Types of Machine Learning Systems

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

Supervised learning involves training a model with labeled data, providing the desired output for each input, so that the model can learn to accurately predict future outputs.

Self-supervised learning involves training a model to predict some parts of data from other parts, such as predicting missing pixels in an image, which can be used to learn useful representations of the data

Unsupervised algorithms detect hidden patterns without human intervention (hence, the term "unsupervised").

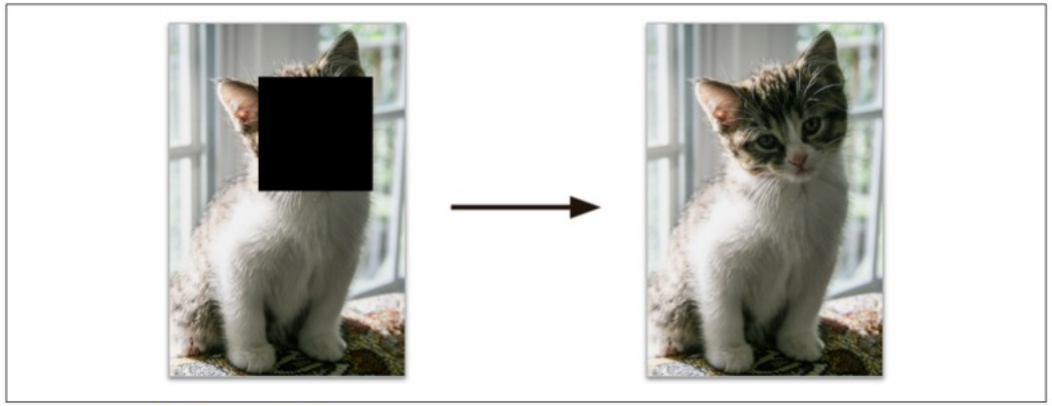


Figure 1-12. Self-supervised learning example: input (left) and target (right)

Types of Machine Learning Systems

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

Semi-supervised learning is a combination of supervised and unsupervised learning. In this approach, a small amount of labeled data is used along with a large amount of unlabeled data, which combines the benefits of both unsupervised and supervised learning while avoiding the challenges associated with finding a large amount of labeled data.

Reinforcement learning involves training a model to make decisions in an environment based on a reward system, where the model learns to maximize its reward over time by taking actions that lead to positive outcomes.

The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return (or penalties in the form of negative rewards, as shown in Figure 1-13).

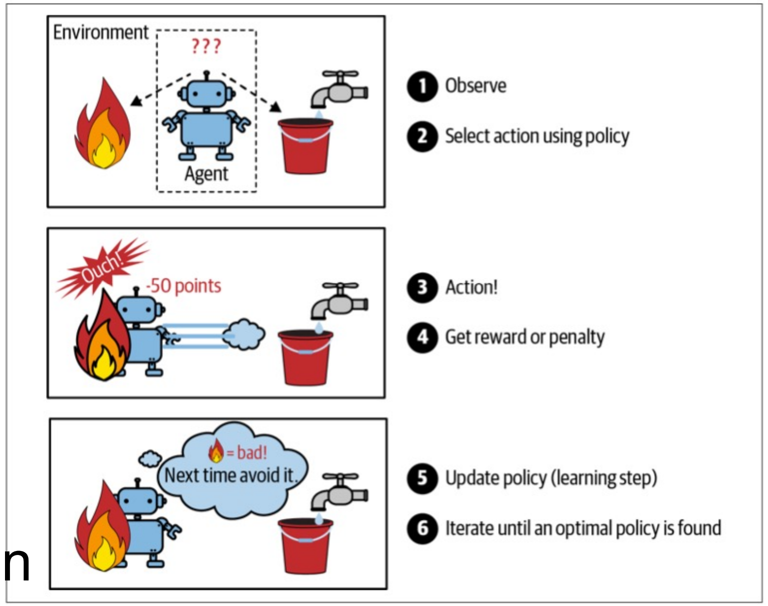

Environment
???
Agent
1 Observe
2 Select action using policy
Ouch!
-50 points
3 Action!
4 Get reward or penalty
= bad!
Next time avoid it.
5 Update policy (learning step)
6 Iterate until an optimal policy is found

Figure 1-13. Reinforcement learning

Types of Machine Learning Systems

Q: What is the difference between batch and online learning?

- In batch learning, the model is trained on the entire dataset at once, while in online learning, the model is trained on small batches of data sequentially.
- Frequency of parameter updates: In batch learning, the model makes predictions on the entire dataset before any updates are made, while in online learning, the model can update its parameters after processing each batch
- Adaptability to changing data: Online learning is more adaptable to changing data as it can quickly update the model parameters with each new batch of data, while batch learning requires retraining the model from scratch if the dataset changes.
- Memory requirements: Batch learning requires more memory to store the entire dataset

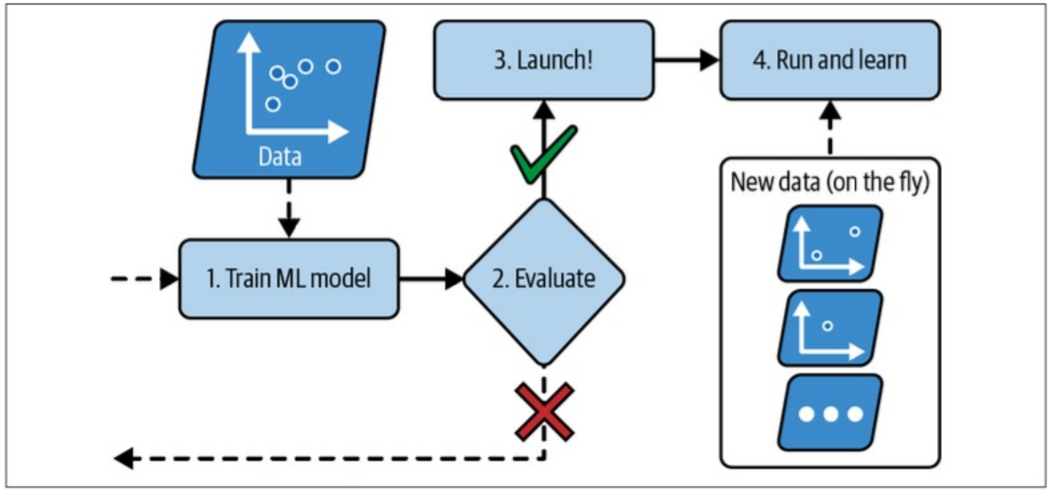


Figure 1-14. In online learning, a model is trained and launched into production, and then it keeps learning as new data comes in

Types of Machine Learning Systems

Q: What is the disadvantage of batch learning? Do you already know of models that definitely do batch learning?

- The main disadvantage of batch learning is its high memory requirement to store the entire dataset.
- Batch learning can be slower than online learning for large datasets, as the model needs to process the entire dataset before making any updates.
- Models that are used batch learning Linear Regression, Logistic Regression, Support Vector Machines (SVMs), and Decision Trees

Types of Machine Learning Systems

Q: What does the term “out-of-core learning” mean? What are challenges in online learning?

Out-of-core learning refers to machine **learning algorithms working with data that cannot fit into a single machine's memory but can easily fit into some data storage, such as a local hard disk or web repository**.

Challenges:

- Stability of the training process.
- Dealing with concept drift

An example of online learning with concept drift is shown in the following picture from the book. The dataset is a time series of housing prices, and the task is to predict the price of the houses.

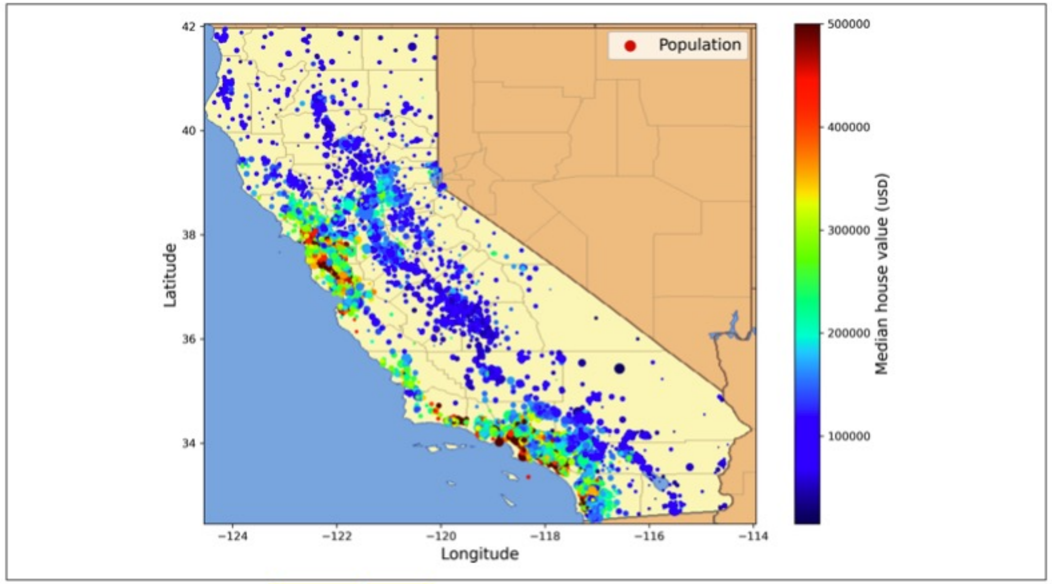


Figure 2-1. California housing prices

The main challenges in Machine Learning

Q: Briefly name all the challenges we encounter in Machine Learning

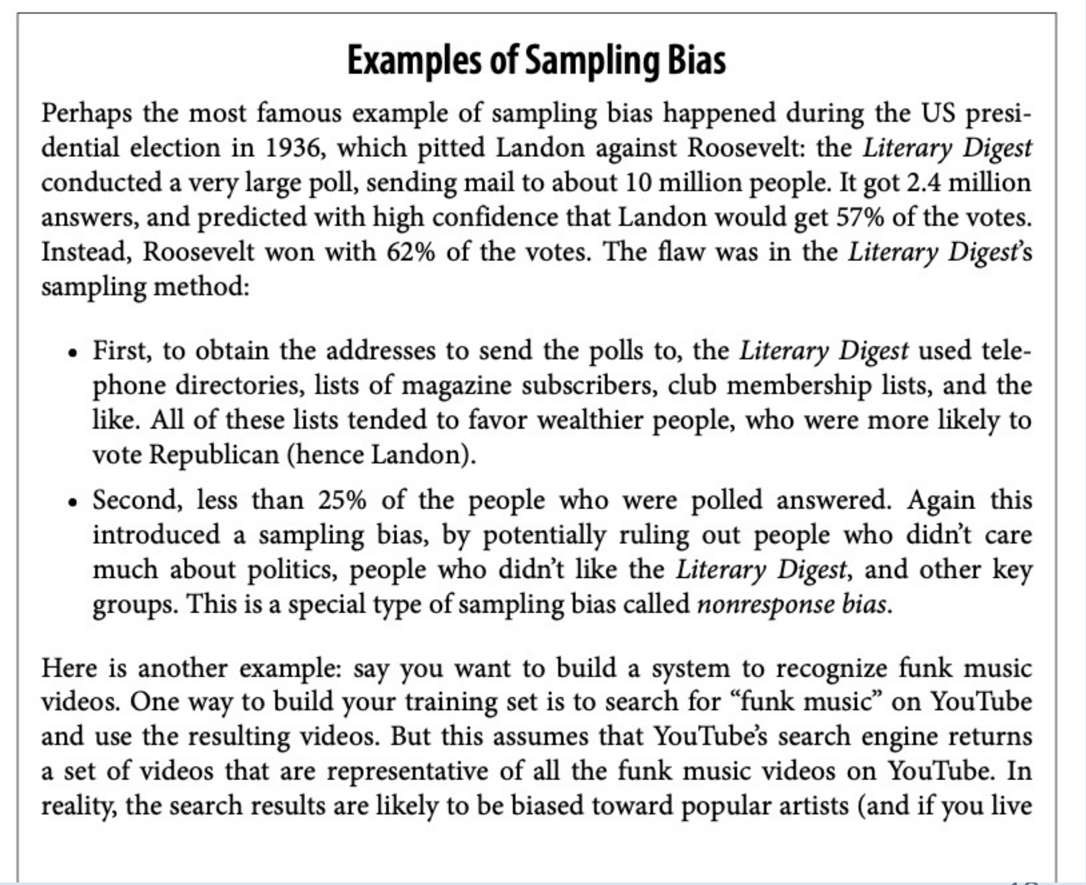
- Insufficient quantity of training data
- Non-representative training data
- Poor-quality data
- Irrelevant features
- Overfitting
- Underfitting
- Computational complexity
- Interpretability

The main challenges in Machine Learning

Q: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 the variability or randomness that occurs when we take a sample from a larger population. This can happen because the sample is too small or because there is natural variation in the population. Due to sampling noise, our model may not make accurate predictions because the data we used to train it may not be representative of the population.

Sampling bias occurs when the sample we collect is not representative of the population we are interested in.



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

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

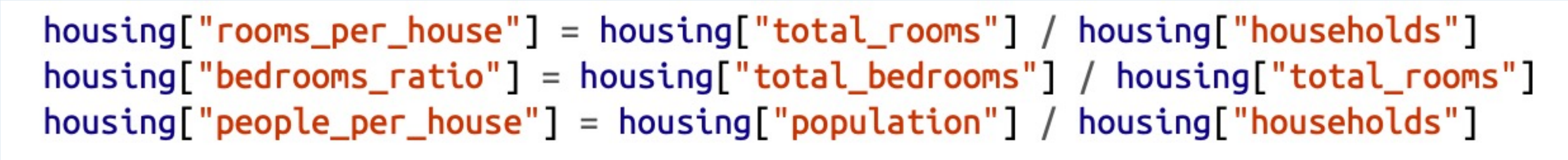
The input variables that we give to our machine learning models are called features.

- Numeric
- Categorical
- Textual

The main challenges in Machine Learning

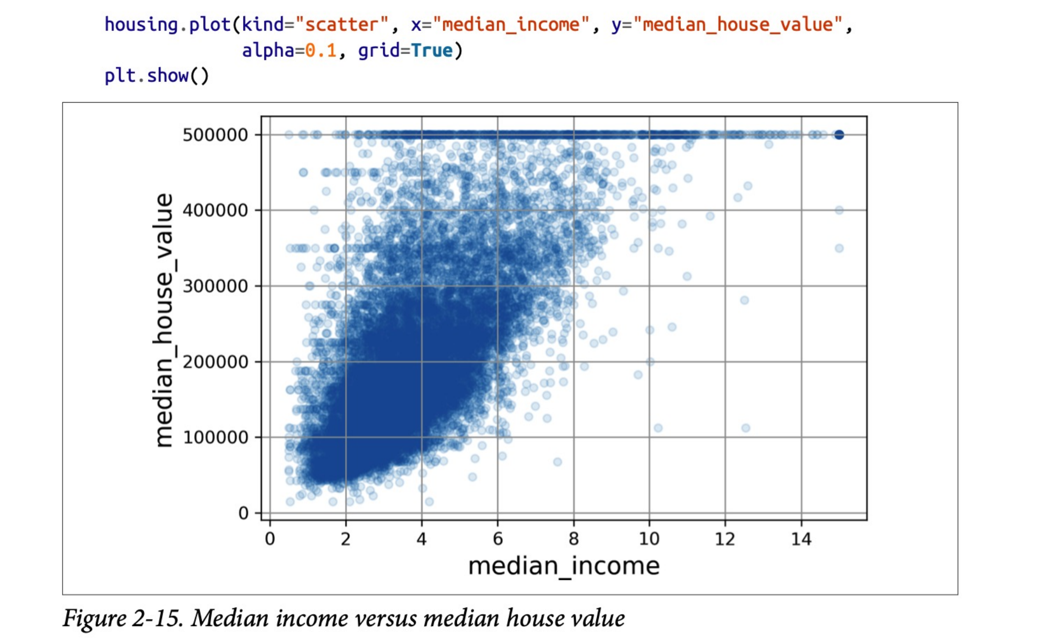
Q:What are features? Use a simple ML model and describe what is meant by feature selection and feature extraction?

In this example, the author extracted a new feature by combining the total number of bedrooms and the total number of rooms in each neighborhood to obtain the bedroom-to-room ratio, which is a measure of how many bedrooms there are per room.



```
housing["rooms_per_house"] = housing["total_rooms"] / housing["households"]
housing["bedrooms_ratio"] = housing["total_bedrooms"] / housing["total_rooms"]
housing["people_per_house"] = housing["population"] / housing["households"]
```

In the sample of the housing price prediction model, the author decided to use only median income as a feature, as it was found to be the most important predictor of housing prices.



```
housing.plot(kind="scatter", x="median_income", y="median_house_value",
             alpha=0.1, grid=True)
plt.show()
```

Figure 2-15. Median income versus median house value

The main challenges in Machine Learning

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

Overfitting occurs when a machine learning model becomes too complex and starts to fit the training data too closely, making it less effective at generalizing to new, unseen data. Having too many features or parameters in comparison to the quantity of training data, adopting a model that is too sophisticated, or training for an excessive length of time can all favor overfitting.

There are a number of ways to prevent overfitting, including regularization, which adds a penalty term to the loss function to deter the model from fitting the training data too closely, or lowering the complexity of the model by deleting features. To improve the diversity of the training data, another technique is to employ more training data or to use data augmentation.

The main challenges in Machine Learning

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

Underfitting happens when a machine learning model is too simple and fails to capture the underlying patterns in the data, resulting in poor performance on both the training and test data. Underfitting can be favored by using a model that is too simple, using too few features or parameters, or not training for enough iterations.

To prevent underfitting, several techniques can be used such as increasing the complexity of the model by adding more layers or neurons, increasing the number of features or input data, or training for more iterations. Another technique is to use a more powerful model architecture, such as a deep neural network, or to fine-tune the hyperparameters of the model.

Testing and Validating

Q: What is the motivation of holdout validation?

The hold-out method is used **to check how well a machine learning model will perform on the new data**.

Testing and Validating

Q: Using Figure 1-25, try to explain what is the test set, the training set, and the validation set?

Using Figure 1-25 the dataset is split into three sets:

- the Training set
- the Dev set (validation)
- the Test set

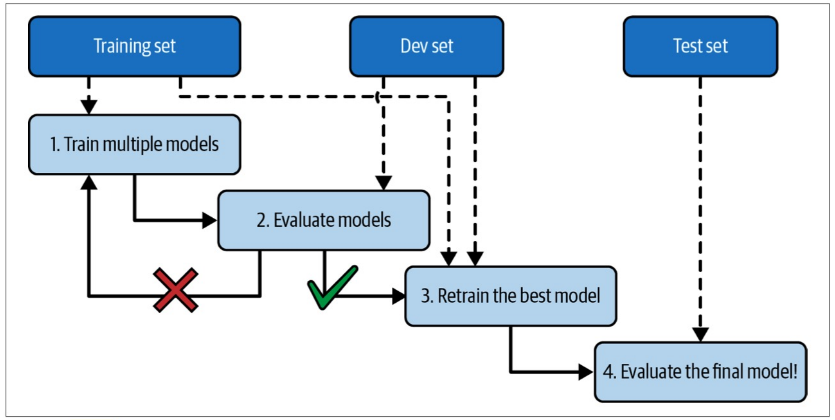


Figure 1-25. Model selection using holdout validation

The training set is used to train the model, and the validation set is used to tune the hyperparameters of the model.

The validation set is a smaller subset of the data that is used to evaluate the performance of the model on new data that it has not seen before.

The test set is used to evaluate the final performance of the model after all hyperparameters have been tuned.

Thank you

Igor Gatchin

Igor.Gatchin@student.uibk.ac.at