5 Model-Agnostic Methods

Separating the explanations from the machine learning model (= model-agnostic interpretation methods) has some advantages (Ribeiro, Singh, and Guestrin 2016[1]). The great advantage of model-agnostic interpretation methods over model-specific ones is their flexibility. Machine learning developers are free to use any machine learning model they like when the interpretation methods can be applied to any model. Anything that builds on an interpretation of a machine learning model, such as a graphic or user interface, also becomes independent of the underlying machine learning model. Typically, not just one, but many types of machine learning models are evaluated to solve a task, and when comparing models in terms of interpretability, it is easier to work with model-agnostic explanations, because the same method can be used for any type of model.

An alternative to model-agnostic interpretation methods is to use only interpretable models, which often has the big disadvantage that predictive performance is lost compared to other machine learning models and you limit yourself to one type of model. The other alternative is to use model-specific interpretation methods. The disadvantage of this is that it also binds you to one model type and it will be difficult to switch to something else.

Desirable aspects of a model-agnostic explanation system are (Ribeiro, Singh, and Guestrin 2016):

- **Model flexibility:** The interpretation method can work with any machine learning model, such as random forests and deep neural networks.
- **Explanation flexibility:** You are not limited to a certain form of explanation. In some cases it might be useful to have a linear formula, in other cases a graphic with feature importances.
- **Representation flexibility:** The explanation system should be able to use a different feature representation as the model being explained. For a text classifier that uses abstract word embedding vectors, it might be preferable to use the presence of individual words for the explanation.

The bigger picture

[1]Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning." ICML Workshop on Human Interpretability in Machine Learning. (2016).

Let us take a high level look at model-agnostic interpretability. We capture the world by collecting data, and abstract it further by learning to predict the data (for the task) with a machine learning model. Interpretability is just another layer on top that helps humans understand.

The lowest layer is the **World**. This could literally be nature itself, like the biology of the human body and how it reacts to medication, but also more abstract things like the real estate market. The World layer contains everything that can be observed and is of interest. Ultimately, we want to learn something about the World and interact with it.

The second layer is the **Data** layer. We have to digitize the World in order to make it processable for computers and also to store information. The Data layer contains anything from images, texts, tabular data and so on.

By fitting machine learning models based on the Data layer, we get the **Black Box Model** layer. Machine learning algorithms learn with data from the real world to make predictions or find structures.

Above the Black Box Model layer is the **Interpretability Methods** layer, which helps us deal with the opacity of machine learning models. What were the most important features for a particular diagnosis? Why was a financial transaction classified as fraud?

The last layer is occupied by a **Human**. Look! This one waves to you because you are reading this book and helping to provide better explanations for black box models! Humans are ultimately the consumers of the explanations.

This multi-layered abstraction also helps to understand the differences in approaches between statisticians and machine learning practitioners. Statisticians deal with the Data layer, such as planning clinical trials or designing surveys. They skip the Black Box Model layer and go right to the Interpretability Methods layer. Machine learning specialists also deal with the Data layer, such as collecting labeled samples of skin cancer images or crawling Wikipedia. Then they train a black box machine learning model. The Interpretability Methods layer is skipped and humans directly deal with the black box model predictions. It's great that interpretable machine learning fuses the work of statisticians and machine learning specialists.

Of course this graphic does not capture everything: Data could come from simulations. Black box models also output predictions that might not even reach humans, but only supply other machines, and so on. But overall it is a useful abstraction to understand how interpretability becomes this new layer on top of machine learning models.

Model-agnostic interpretation methods can be further distinguished into local and global methods. The book is also organized according to this distinction. Global methods describe how features affect the prediction **on average**. In contrast, local methods aim to explain **individual predictions**.

模型无关的解释方法分为：
全局：feature 平均上如何影响预测
局部：旨在解释个别预测
What is 'individual predictions'?

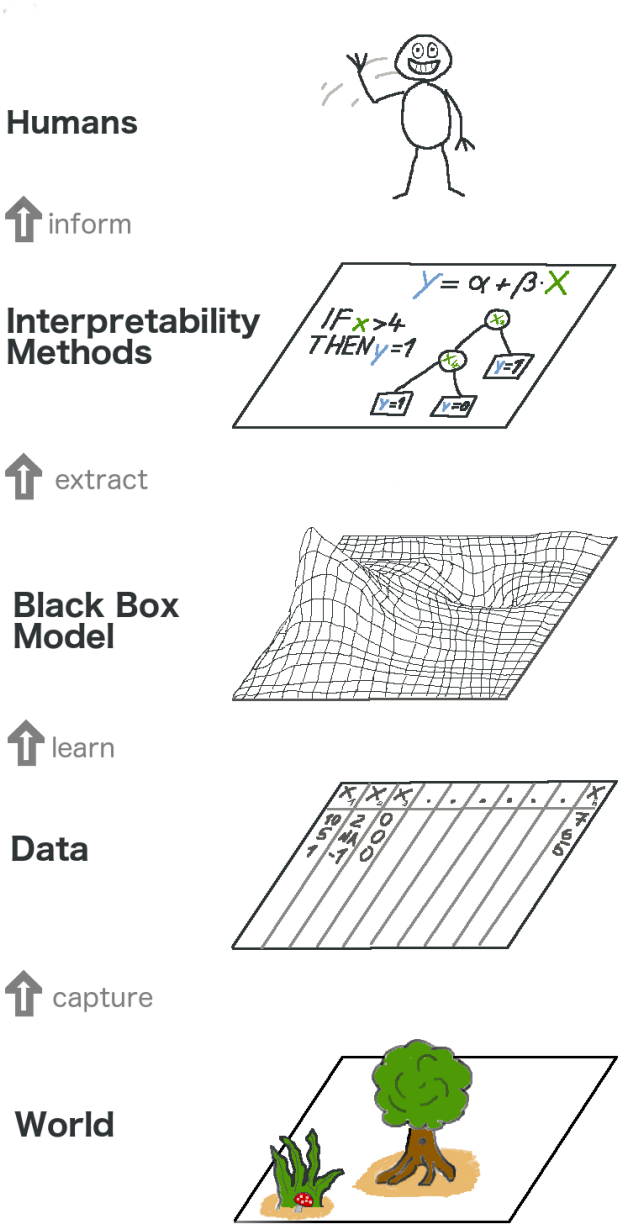

Figure 5.1: The big picture of explainable machine learning. The real world goes through many layers before it reaches the human in the form of explanations.

6 Example-Based Explanations

基于实例的解释

Example-based explanation methods select particular instances of the dataset to explain the behavior of machine learning models or to explain the underlying data distribution.

Summaries of features:
① feature importance
② partial dependence

Example-based explanations are mostly model-agnostic, because they make any machine learning model more interpretable. The difference to model-agnostic methods is that the example-based methods explain a model by selecting instances of the dataset and not by creating summaries of features (such as feature importance or partial dependence). Example-based explanations only make sense if we can represent an instance of the data in a humanly understandable way. This works well for images, because we can view them directly. In general, example-based methods work well if the feature values of an instance carry more context, meaning the data has a structure, like images or texts do. It is more challenging to represent tabular data in a meaningful way, because an instance can consist of hundreds or thousands of (less structured) features. Listing all feature values to describe an instance is usually not useful. It works well if there are only a handful of features or if we have a way to summarize an instance.

tabular 很难用 example-based 解释，尤其是有很多 features 时

Example-based explanations help humans construct mental models of the machine learning model and the data the machine learning model has been trained on. It especially helps to understand complex data distributions. But what do I mean by example-based explanations? We often use them in our jobs and daily lives. Let us start with some examples[1].

A physician sees a patient with an unusual cough and a mild fever. The patient's symptoms remind her of another patient she had years ago with similar symptoms. She suspects that her current patient could have the same disease and she takes a blood sample to test for this specific disease.

A data scientist works on a new project for one of his clients: Analysis of the risk factors that lead to the failure of production machines for keyboards. The data scientist remembers a similar project he worked on and reuses parts of the code from the old project because he thinks the client wants the same analysis.

A kitten sits on the window ledge of a burning and uninhabited house. The fire department has already arrived and one of the firefighters ponders for a second whether he can risk

[1] Aamodt, Agnar, and Enric Plaza. "Case-based reasoning: Foundational issues, methodological variations, and system approaches." AI communications 7.1 (1994): 39-59.

going into the building to save the kitten. He remembers similar cases in his life as a firefighter: Old wooden houses that have been burning slowly for some time were often unstable and eventually collapsed. Because of the similarity of this case, he decides not to enter, because the risk of the house collapsing is too great. Fortunately, the kitty jumps out of the window, lands safely and nobody is harmed in the fire. Happy end.

These stories illustrate how we humans think in examples or analogies. The blueprint of example-based explanations is: Thing B is similar to thing A and A caused Y, so I predict that B will cause Y as well. Implicitly, some machine learning approaches work example-based. Decision trees partition the data into nodes based on the similarities of the data points in the features that are important for predicting the target. A decision tree gets the prediction for a new data instance by finding the instances that are similar (= in the same terminal node) and returning the average of the outcomes of those instances as the prediction. The k-nearest neighbors (knn) method works explicitly with example-based predictions. For a new instance, a knn model locates the k-nearest neighbors (e.g. the k=3 closest instances) and returns the average of the outcomes of those neighbors as a prediction. The prediction of a knn can be explained by returning the k neighbors, which – again – is only meaningful if we have a good way to represent a single instance.

The following interpretation methods are all example-based:

- Counterfactual explanations tell us how an instance has to change to significantly change its prediction. By creating counterfactual instances, we learn about how the model makes its predictions and can explain individual predictions.
- Adversarial examples are counterfactuals used to fool machine learning models. The emphasis is on flipping the prediction and not explaining it.
- Prototypes are a selection of representative instances from the data and criticisms are instances that are not well represented by those prototypes. [2]
- Influential instances are the training data points that were the most influential for the parameters of a prediction model or the predictions themselves. Identifying and analysing influential instances helps to find problems with the data, debug the model and understand the model's behavior better.
- k-nearest neighbors model: An (interpretable) machine learning model based on examples.

[2]Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).

决策树

k-nn

反事实样本：找出让 model 反应大的例子

对抗样本：欺骗 model

有影响力的样本

k-nn 就是基于样本的