7 Global Model-Agnostic Methods

Global methods describe the average behavior of a machine learning model. The counterpart to global methods are local methods. Global methods are often expressed as expected values based on the distribution of the data. For example, the partial dependence plot, a feature effect plot, is the expected prediction when all other features are marginalized out. Since global interpretation methods describe average behavior, they are particularly useful when the modeler wants to understand the general mechanisms in the data or debug a model.

In this book, you will learn about the following model-agnostic global interpretation techniques:

- The partial dependence plot is a feature effect method.
- Accumulated local effect plots is another feature effect method that works when features are dependent.
- Feature interaction (H-statistic) quantifies to what extent the prediction is the result of joint effects of the features.
- Functional decomposition is a central idea of interpretability and a technique that decomposes the complex prediction function into smaller parts.
- Permutation feature importance measures the importance of a feature as an increase in loss when the feature is permuted.
- Global surrogate models replaces the original model with a simpler model for interpretation.
- Prototypes and criticisms are representative data point of a distribution and can be used to enhance interpretability.

average behavior of the model

marginalized 排除；忽略

PDP

功能拆分

排列特征的重要性

全局代理 model

有代表性的样本

7.1 Partial Dependence Plot (PDP)

The partial dependence plot (short PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model (J. H. Friedman 2001[1]). A partial dependence plot can show whether the relationship between the target and a feature is linear, monotonic or more complex. For example, when applied to a linear regression model, partial dependence plots always show a linear relationship.

The partial dependence function for regression is defined as:

$$\hat{f}_S(x_S) = E_{X_C}\left[\hat{f}(x_S, X_C)\right] = \int \hat{f}(x_S, X_C) d\mathbb{P}(X_C)$$

The x_S are the features for which the partial dependence function should be plotted and X_C are the other features used in the machine learning model $\hat{f}$, which are here treated as random variables. Usually, there are only one or two features in the set S. The feature(s) in S are those for which we want to know the effect on the prediction. The feature vectors x_S and x_C combined make up the total feature space x. Partial dependence works by marginalizing the machine learning model output over the distribution of the features in set C, so that the function shows the relationship between the features in set S we are interested in and the predicted outcome. By marginalizing over the other features, we get a function that depends only on features in S, interactions with other features included.

The partial function $\hat{f}_S$ is estimated by calculating averages in the training data, also known as Monte Carlo method:

$$\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

The partial function tells us for given value(s) of features S what the average marginal effect on the prediction is. In this formula, $x_C^{(i)}$ are actual feature values from the dataset for the features in which we are not interested, and n is the number of instances in the dataset. An assumption of the PDP is that the features in C are not correlated with the features in S. If this assumption is violated, the averages calculated for the partial dependence plot will include data points that are very unlikely or even impossible (see disadvantages).

For classification where the machine learning model outputs probabilities, the partial dependence plot displays the probability for a certain class given different values for feature(s) in S. An easy way to deal with multiple classes is to draw one line or plot per class.

[1] Friedman, Jerome H. "Greedy function approximation: A gradient boosting machine." Annals of statistics (2001): 1189-1232.

marginal effect 边缘效应
只变一个 feature 看它怎么影响 model 预测
PDP 只变一到二个 features

x_S 是要测的 features
X_C 是其它 features 是随机值
x_S 和 x_C 组成了所有的 features

PD 是通过将集合 C 中的 features 分布边缘化而起作用的

$\hat{f}_S$ 是训练集中计算出来的平均值

PDP 的求法

$x_C^{(i)}$ 是实际值
n 是样本数

The partial dependence plot is a global method: The method considers all instances and gives a statement about the global relationship of a feature with the predicted outcome.

Categorical features

So far, we have only considered numerical features. For categorical features, the partial dependence is very easy to calculate. For each of the categories, we get a PDP estimate by forcing all data instances to have the same category. For example, if we look at the bike rental dataset and are interested in the partial dependence plot for the season, we get four numbers, one for each season. To compute the value for "summer", we replace the season of all data instances with "summer" and average the predictions.

7.1.1 PDP-based Feature Importance

Greenwell et al. (2018) [2] proposed a simple partial dependence-based feature importance measure. The basic motivation is that a flat PDP indicates that the feature is not important, and the more the PDP varies, the more important the feature is. For numerical features, importance is defined as the deviation of each unique feature value from the average curve:

$$I(x_S) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(\hat{f}_S(x_S^{(k)}) - \frac{1}{K}\sum_{k=1}^{K}\hat{f}_S(x_S^{(k)}))^2}$$

Note that here the $x_S^{(k)}$ are the K unique values of feature the X_S. For categorical features we have:

$$I(x_S) = (max_k(\hat{f}_S(x_S^{(k)})) - min_k(\hat{f}_S(x_S^{(k)})))/4$$

This is the range of the PDP values for the unique categories divided by four. This strange way of calculating the deviation is called the range rule. It helps to get a rough estimate for the deviation when you only know the range. And the denominator four comes from the standard normal distribution: In the normal distribution, 95% of the data are minus two and plus two standard deviations around the mean. So the range divided by four gives a rough estimate that probably underestimates the actual variance.

This PDP-based feature importance should be interpreted with care. It captures only the main effect of the feature and ignores possible feature interactions. A feature could be very important based on other methods such as permutation feature importance, but the PDP could be flat as the feature affects the prediction mainly through interactions with

[2] Greenwell, Brandon M., Bradley C. Boehmke, and Andrew J. McCarthy. "A simple and effective model-based variable importance measure." arXiv preprint arXiv:1805.04755 (2018).

PDP越变化大，feature的重要性越大。

数值型

$x_S^{(k)}$ 中的K代表这个数值feature的值的权举数量

绝对型feature

4来自于正态分布

denominator 分母；统计基准

PDP的解释要小心，它忽略了feature之间的交互

缺点1：

feature可能因为其它feature的交互而导致PDP很平

other features. Another drawback of this measure is that it is defined over the unique values. A unique feature value with just one instance is given the same weight in the importance computation as a value with many instances.

7.1.2 Examples

In practice, the set of features S usually only contains one feature or a maximum of two, because one feature produces 2D plots and two features produce 3D plots. Everything beyond that is quite tricky. Even 3D on a 2D paper or monitor is already challenging.

Let us return to the regression example, in which we predict the number of bikes that will be rented on a given day. First we fit a machine learning model, then we analyze the partial dependencies. In this case, we have fitted a random forest to predict the number of bicycles and use the partial dependence plot to visualize the relationships the model has learned. The influence of the weather features on the predicted bike counts is visualized in the following figure.

For warm but not too hot weather, the model predicts on average a high number of rented bicycles. Potential bikers are increasingly inhibited in renting a bike when humidity exceeds 60%. In addition, the more wind the fewer people like to cycle, which makes sense. Interestingly, the predicted number of bike rentals does not fall when wind speed increases from 25 to 35 km/h, but there is not much training data, so the machine learning model could probably not learn a meaningful prediction for this range. At least intuitively, I would expect the number of bicycles to decrease with increasing wind speed, especially when the wind speed is very high.

To illustrate a partial dependence plot with a categorical feature, we examine the effect of the season feature on the predicted bike rentals.

We also compute the partial dependence for cervical cancer classification. This time we fit a random forest to predict whether a woman might get cervical cancer based on risk factors. We compute and visualize the partial dependence of the cancer probability on different features for the random forest:

We can also visualize the partial dependence of two features at once:

7.1.3 Advantages

The computation of partial dependence plots is **intuitive**: The partial dependence function at a particular feature value represents the average prediction if we force all data points to assume that feature value. In my experience, lay people usually understand the idea of PDPs quickly.

缺点2：
$I(x_S)$是根据独特值来测量的，如果仅有一个样本量生了独特值而影响到了PDP，那它的影响力和大多数样本一样大。

Fig 7.1

Fig 7.2

Fig 7.3

Fig 7.4

优点：直观易懂。

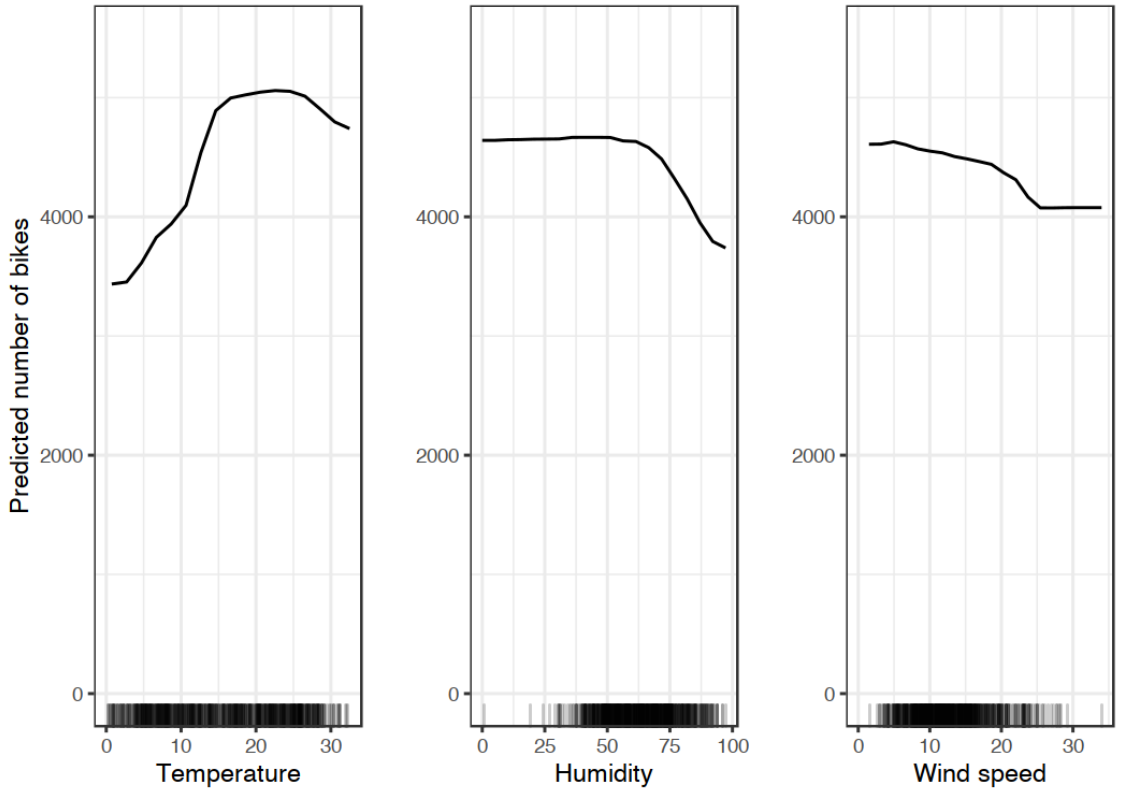

Figure 7.1: PDPs for the bicycle count prediction model and temperature, humidity and wind speed. The largest differences can be seen in the temperature. The hotter, the more bikes are rented. This trend goes up to 20 degrees Celsius, then flattens and drops slightly at 30. Marks on the x-axis indicate the data distribution.

If the feature for which you computed the PDP is not correlated with the other features, then the PDPs perfectly represent how the feature influences the prediction on average. In the uncorrelated case, the **interpretation is clear**: The partial dependence plot shows how the average prediction in your dataset changes when the j-th feature is changed. It is more complicated when features are correlated, see also disadvantages.

Partial dependence plots are **easy to implement**.

The calculation for the partial dependence plots has a **causal interpretation**. We intervene on a feature and measure the changes in the predictions. In doing so, we analyze the causal relationship between the feature and the prediction.[3] The relationship is causal

[3]Zhao, Qingyuan, and Trevor Hastie. "Causal interpretations of black-box models." Journal of Business & Economic Statistics, to appear. (2017).

PDP 的 feature 和其它 feature 没有联系(交互)的话，那 PDP 的解释是完美的。

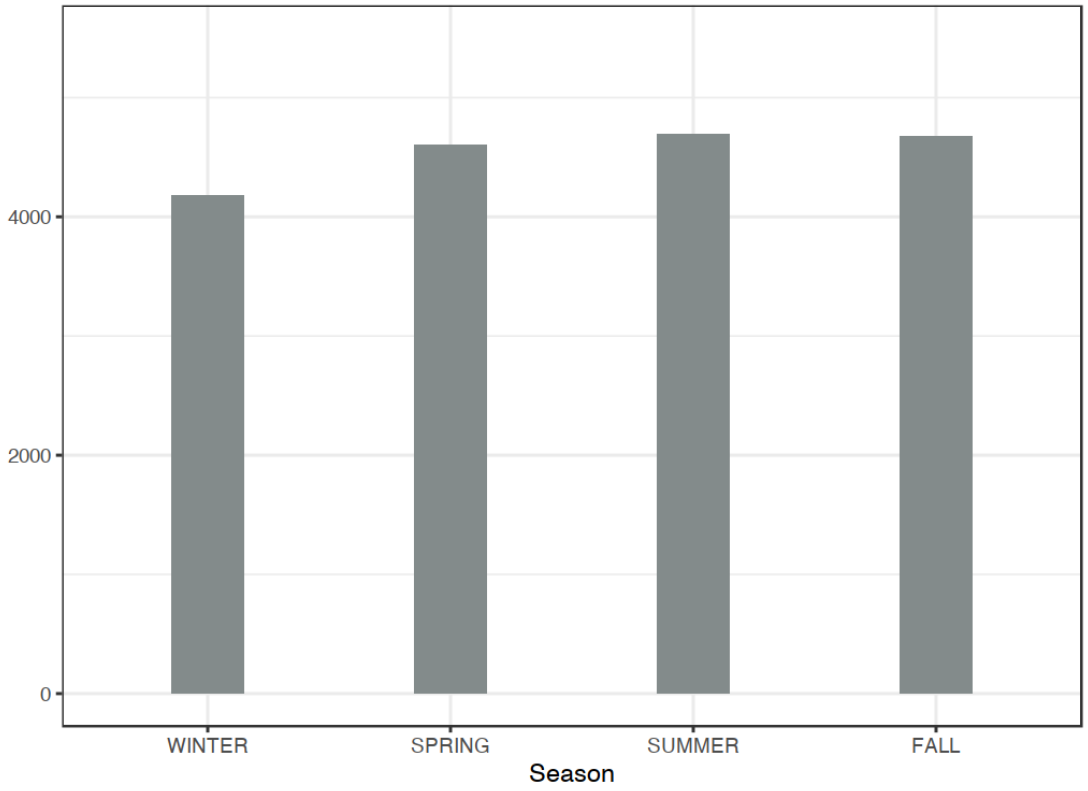

Figure 7.2: PDPs for the bike count prediction model and the season. Unexpectedly all seasons show similar effect on the model predictions, only for winter the model predicts fewer bicycle rentals.

for the model – because we explicitly model the outcome as a function of the features – but not necessarily for the real world!

7.1.4 Disadvantages

The realistic **maximum number of features** in a partial dependence function is two. This is not the fault of PDPs, but of the 2-dimensional representation (paper or screen) and also of our inability to imagine more than 3 dimensions.

Some PD plots do not show the **feature distribution**. Omitting the distribution can be misleading, because you might overinterpret regions with almost no data. This problem is easily solved by showing a rug (indicators for data points on the x-axis) or a histogram.

The **assumption of independence** is the biggest issue with PD plots. It is assumed that the feature(s) for which the partial dependence is computed are not correlated with other

→ PDP for categorical features

缺点：

① 最多同时测两个 features 的 PDP，因为我们只能理解3D的图

② PDP不展示features的分布

③ PDP的解释只成立在被测 feature(s) 和其它 features 之间没有联系的情况下

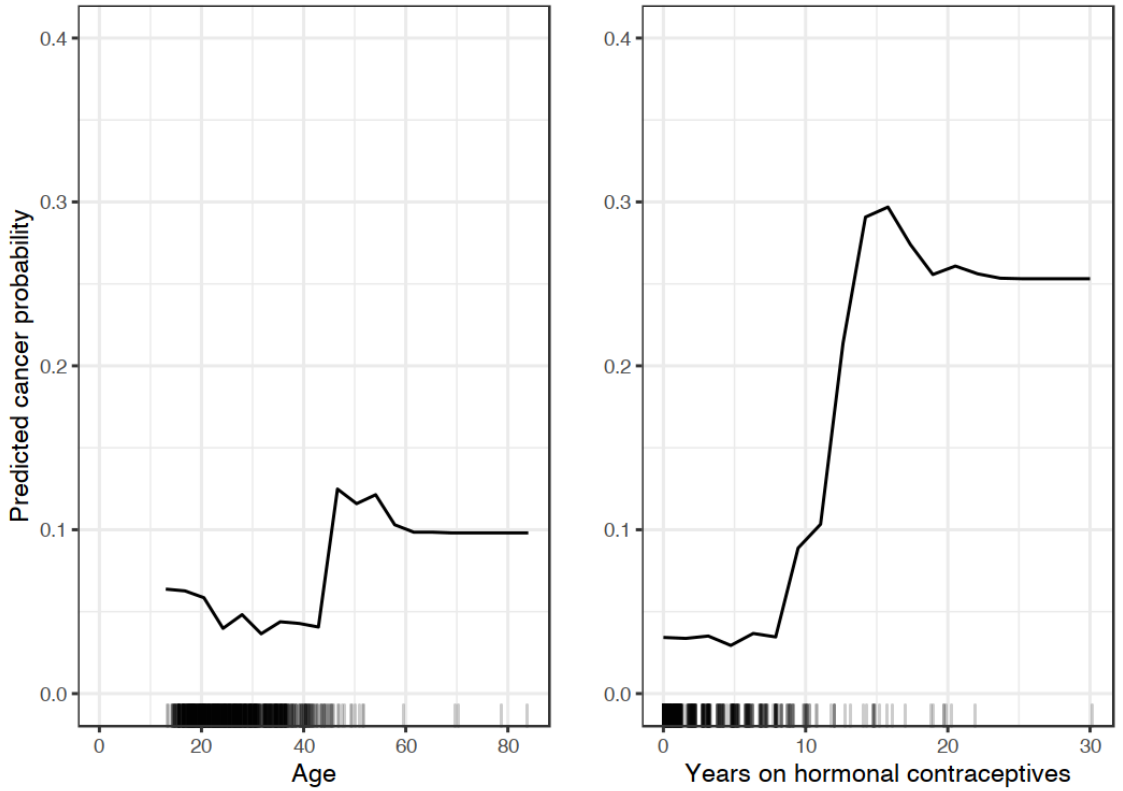

Figure 7.3: PDPs of cancer probability based on age and years with hormonal contraceptives. For age, the PDP shows that the probability is low until 40 and increases after. The more years on hormonal contraceptives the higher the predicted cancer risk, especially after 10 years. For both features not many data points with large values were available, so the PD estimates are less reliable in those regions.

features. For example, suppose you want to predict how fast a person walks, given the person's weight and height. For the partial dependence of one of the features, e.g. height, we assume that the other features (weight) are not correlated with height, which is obviously a false assumption. For the computation of the PDP at a certain height (e.g. 200 cm), we average over the marginal distribution of weight, which might include a weight below 50 kg, which is unrealistic for a 2 meter person. In other words: When the features are correlated, we create new data points in areas of the feature distribution where the actual probability is very low (for example it is unlikely that someone is 2 meters tall but weighs less than 50 kg). One solution to this problem is Accumulated Local Effect plots or short ALE plots that work with the conditional instead of the marginal distribution.

Heterogeneous effects might be hidden because PD plots only show the average

缺点3的解决方案：
使用适用于条件分布而不是
边缘分布的 ALE plots

heterogeneous effect
非均匀效应
中文书翻译为"异质效应"

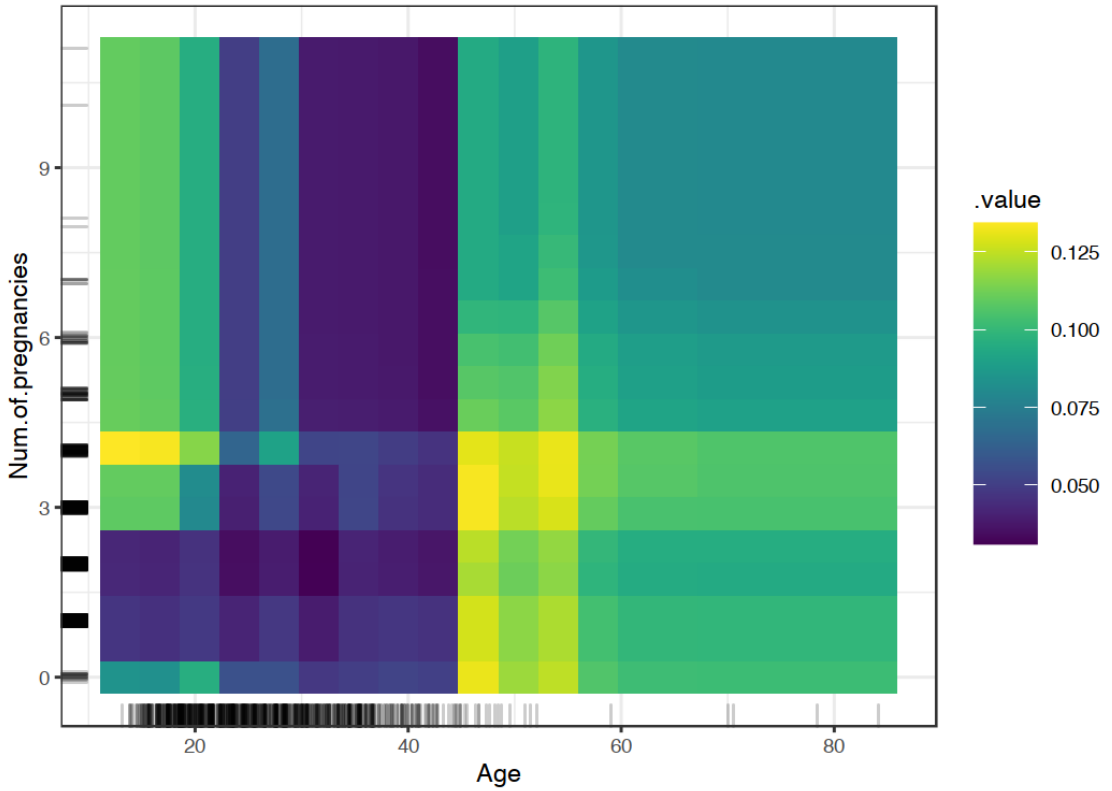

Figure 7.4: PDP of cancer probability and the interaction of age and number of pregnancies. The plot shows the increase in cancer probability at 45. For ages below 25, women who had 1 or 2 pregnancies have a lower predicted cancer risk, compared with women who had 0 or more than 2 pregnancies. But be careful when drawing conclusions: This might just be a correlation and not causal!

marginal effects. Suppose that for a feature half your data points have a positive association with the prediction – the larger the feature value the larger the prediction – and the other half has a negative association – the smaller the feature value the larger the prediction. The PD curve could be a horizontal line, since the effects of both halves of the dataset could cancel each other out. You then conclude that the feature has no effect on the prediction. By plotting the individual conditional expectation curves instead of the aggregated line, we can uncover heterogeneous effects.

7.1.5 Software and Alternatives

There are a number of R packages that implement PDPs. I used the `iml` package for the examples, but there is also `pdp` or `DALEX`. In Python, partial dependence plots are built

3D Plot for PDP of 2features

It can reflect the interaction between two features

假设对于一个feature，数据点中一半与预测正相关，另一半负相关

PD 曲线可能是一条水平线，因为正负相关抵消了，因此可能会作出错误的解释。

ICE可以解决这个问题

into `scikit-learn` and you can use `PDPBox`.

Alternatives to PDPs presented in this book are ALE plots and ICE curves.