4 Interpretable Models

The easiest way to achieve interpretability is to use only a subset of algorithms that create interpretable models. Linear regression, logistic regression and the decision tree are commonly used interpretable models.

In the following chapters we will talk about these models. Not in detail, only the basics, because there is already a ton of books, videos, tutorials, papers and more material available. We will focus on how to interpret the models. The book discusses linear regression, logistic regression, other linear regression extensions, decision trees, decision rules and the RuleFit algorithm in more detail. It also lists other interpretable models.

All interpretable models explained in this book are interpretable on a modular level, with the exception of the k-nearest neighbors method. The following table gives an overview of the interpretable model types and their properties. A model is linear if the association between features and target is modelled linearly. A model with monotonicity constraints ensures that the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature: An increase in the feature value either always leads to an increase or always to a decrease in the target outcome. Monotonicity is useful for the interpretation of a model because it makes it easier to understand a relationship. Some models can automatically include interactions between features to predict the target outcome. You can include interactions in any type of model by manually creating interaction features. Interactions can improve predictive performance, but too many or too complex interactions can hurt interpretability. Some models handle only regression, some only classification, and still others both.

单调性

What is interactions ?

From this table, you can select a suitable interpretable model for your task, either regression (regr) or classification (class):

Algorithm	Linear	Monotone	Interaction	Task
Linear regression	Yes	Yes	No	regr
Logistic regression	No	Yes	No	class
Decision trees	No	Some	Yes	class,regr
RuleFit	Yes	No	Yes	class,regr
Naive Bayes	No	Yes	No	class
k-nearest neighbors	No	No	No	class,regr

You could argue that both logistic regression and Naive Bayes allow linear explanations. However, this is only true for the logarithm of the target: Increasing a feature by one point increases the **logarithm** of the target probability by a certain amount assuming all other features remain the same.

logarithm 对数

当其它features不变的时候，提升一个feature一个点，使目标概率的对数提升一定量

4.1 Linear Regression

A linear regression model predicts the target as a weighted sum of the feature inputs. The linearity of the learned relationship makes the interpretation easy. Linear regression models have long been used by statisticians, computer scientists and other people who tackle quantitative problems.

weighted sum 加权和

Linear models can be used to model the dependence of a regression target y on some features x. The learned relationships are linear and can be written for a single instance i as follows:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

coefficient 系数

The predicted outcome of an instance is a weighted sum of its p features. The betas (β_j) represent the learned feature weights or coefficients. The first weight in the sum (β_0) is called the intercept and is not multiplied with a feature. The epsilon (ϵ) is the error we still make, i.e. the difference between the prediction and the actual outcome. These errors are assumed to follow a Gaussian distribution, which means that we make errors in both negative and positive directions and make many small errors and few large errors.

error 误差

gaussian 高斯分布

Various methods can be used to estimate the optimal weight. The ordinary least squares method is usually used to find the weights that minimize the squared differences between the actual and the estimated outcomes:

least squares method 最小二乘法

$$\hat{\beta} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2$$

We will not discuss in detail how the optimal weights can be found, but if you are interested, you can read chapter 3.2 of the book "The Elements of Statistical Learning" (Friedman, Hastie and Tibshirani 2009)[1] or one of the other online resources on linear regression models.

The biggest advantage of linear regression models is linearity: It makes the estimation procedure simple and, most importantly, these linear equations have an easy to understand interpretation on a modular level (i.e. the weights). This is one of the main reasons why the linear model and all similar models are so widespread in academic fields such as medicine, sociology, psychology, and many other quantitative research fields. For example, in the medical field, it is not only important to predict the clinical outcome of a patient, but also to quantify the influence of the drug and at the same time take sex, age, and other features into account in an interpretable way.

⇒ 线性model 最容易理解和解释（在模块层面）

[1] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "The elements of statistical learning". www.web.stanford.edu/~hastie/ElemStatLearn/ (2009).

Estimated weights come with confidence intervals. A confidence interval is a range for the weight estimate that covers the "true" weight with a certain confidence. For example, a 95% confidence interval for a weight of 2 could range from 1 to 3. The interpretation of this interval would be: If we repeated the estimation 100 times with newly sampled data, the confidence interval would include the true weight in 95 out of 100 cases, given that the linear regression model is the correct model for the data.

confidence intervals
置信区间

Whether the model is the "correct" model depends on whether the relationships in the data meet certain assumptions, which are linearity, normality, homoscedasticity, independence, fixed features, and absence of multicollinearity.

→ model的正确性取决于数据中的关系是否符合某些假设.

Linearity

The linear regression model forces the prediction to be a linear combination of features, which is both its greatest strength and its greatest limitation. Linearity leads to interpretable models. Linear effects are easy to quantify and describe. They are additive, so it is easy to separate the effects. If you suspect feature interactions or a nonlinear association of a feature with the target value, you can add interaction terms or use regression splines.

线性效应是最容易量化与描述的

Normality

It is assumed that the target outcome given the features follows a normal distribution. If this assumption is violated, the estimated confidence intervals of the feature weights are invalid.

正态的

Homoscedasticity (constant variance)

The variance of the error terms is assumed to be constant over the entire feature space. Suppose you want to predict the value of a house given the living area in square meters. You estimate a linear model that assumes that, regardless of the size of the house, the error around the predicted response has the same variance. This assumption is often violated in reality. In the house example, it is plausible that the variance of error terms around the predicted price is higher for larger houses, since prices are higher and there is more room for price fluctuations. Suppose the average error (difference between predicted and actual price) in your linear regression model is 50,000 Euros. If you assume homoscedasticity, you assume that the average error of 50,000 is the same for houses that cost 1 million and for houses that cost only 40,000. This is unreasonable because it would mean that we can expect negative house prices.

同方差性、
error term 误差项
variance 方差

plausible 似乎是真的

Independence

It is assumed that each instance is independent of any other instance. If you perform repeated measurements, such as multiple blood tests per patient, the data points are not independent. For dependent data you need special linear regression models, such as mixed effect models or GEEs. If you use the "normal" linear regression model, you might draw wrong conclusions from the model.

每个实例都不依赖其它实例 →

Fixed features

The input features are considered "fixed". Fixed means that they are treated as "given constants" and not as statistical variables. This implies that they are free of measurement errors. This is a rather unrealistic assumption. Without that assumption, however, you would have to fit very complex measurement error models that account for the measurement errors of your input features. And usually you do not want to do that.

Absence of multicollinearity

You do not want strongly correlated features, because this messes up the estimation of the weights. In a situation where two features are strongly correlated, it becomes problematic to estimate the weights because the feature effects are additive and it becomes indeterminable to which of the correlated features to attribute the effects.

多重共线性

如果有相关性较强的 feature，
那么它们的 weight 会很难
归因.

4.1.1 Interpretation

The interpretation of a weight in the linear regression model depends on the type of the corresponding feature.

根据 feature 的类型来解释
weight.

- Numerical feature: Increasing the numerical feature by one unit changes the estimated outcome by its weight. An example of a numerical feature is the size of a house.

数值特征：变化数值.

- Binary feature: A feature that takes one of two possible values for each instance. An example is the feature "House comes with a garden". One of the values counts as the reference category (in some programming languages encoded with 0), such as "No garden". Changing the feature from the reference category to the other category changes the estimated outcome by the feature's weight.

二分类特征：变到另一个类型.

- Categorical feature with multiple categories: A feature with a fixed number of possible values. An example is the feature "floor type", with possible categories "carpet", "laminate" and "parquet". A solution to deal with many categories is the one-hot-encoding, meaning that each category has its own binary column. For a categorical feature with L categories, you only need L-1 columns, because the L-th column would have redundant information (e.g. when columns 1 to L-1 all have value 0 for one instance, we know that the categorical feature of this instance takes on category L). The interpretation for each category is then the same as the interpretation for binary features. Some languages, such as R, allow you to encode categorical features in various ways, as described later in this chapter.

分类型（多个分类）：同上

- Intercept β_0: The intercept is the feature weight for the "constant feature", which is always 1 for all instances. Most software packages automatically add this "1"-feature to estimate the intercept. The interpretation is: For an instance with all numerical feature values at zero and the categorical feature values at the reference categories, the model prediction is the intercept weight. The interpretation of the intercept is usually not relevant because instances with all features values at zero

常数：

截距项 β_0

often make no sense. The interpretation is only meaningful when the features have been standardised (mean of zero, standard deviation of one). Then the intercept reflects the predicted outcome of an instance where all features are at their mean value.

The interpretation of the features in the linear regression model can be automated by using following text templates.

Interpretation of a Numerical Feature

An increase of feature x_k by one unit increases the prediction for y by β_k units when all other feature values remain fixed.

Interpretation of a Categorical Feature

Changing feature x_k from the reference category to the other category increases the prediction for y by β_k when all other features remain fixed.

variance 方差；差值

Another important measurement for interpreting linear models is the R-squared measurement. R-squared tells you how much of the total variance of your target outcome is explained by the model. The higher R-squared, the better your model explains the data. The formula for calculating R-squared is:

R-squared 告诉你 model 解释了目标结果总方差的多少.

$$R^2 = 1 - SSE/SST$$

$y^{(i)} - \hat{y}^{(i)}$ 实际值减预测值.

SSE is the squared sum of the error terms:

SSE：误差项平方和

$$SSE = \sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2$$

$\bar{y}$?

SST is the squared sum of the data variance:

SST：数据方差平方和

$$SST = \sum_{i=1}^{n}(y^{(i)} - \bar{y})^2$$

The SSE tells you how much variance remains after fitting the linear model, which is measured by the squared differences between the predicted and actual target values. SST is the total variance of the target outcome. R-squared tells you how much of your variance can be explained by the linear model. R-squared usually ranges between 0 for models where the model does not explain the data at all and 1 for models that explain all of the variance in your data. It is also possible for R-squared to take on a negative value without violating any mathematical rules. This happens when SSE is greater than SST which means that a model does not capture the trend of the data and fits to the data worse than using the mean of the target as the prediction.

R-squared 表示 model 对数据差值的解释程度（0～1）

There is a catch, because R-squared increases with the number of features in the model, even if they do not contain any information about the target value at all. Therefore, it is better to use the adjusted R-squared, which accounts for the number of features used in the model. Its calculation is:

catch n. 隐藏的困难.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where p is the number of features and n the number of instances.

It is not meaningful to interpret a model with very low (adjusted) R-squared, because such a model basically does not explain much of the variance. Any interpretation of the weights would not be meaningful.

Feature Importance

The importance of a feature in a linear regression model can be measured by the absolute value of its t-statistic. The t-statistic is the estimated weight scaled with its standard error.

t-statistic

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

Let us examine what this formula tells us: The importance of a feature increases with increasing weight. This makes sense. The more variance the estimated weight has (= the less certain we are about the correct value), the less important the feature is. This also makes sense.

系数越大，feature越重要.

4.1.2 Example

In this example, we use the linear regression model to predict the number of rented bikes on a particular day, given weather and calendar information. For the interpretation, we examine the estimated regression weights. The features consist of numerical and categorical features. For each feature, the table shows the estimated weight, the standard error of the estimate (SE), and the absolute value of the t-statistic (|t|).