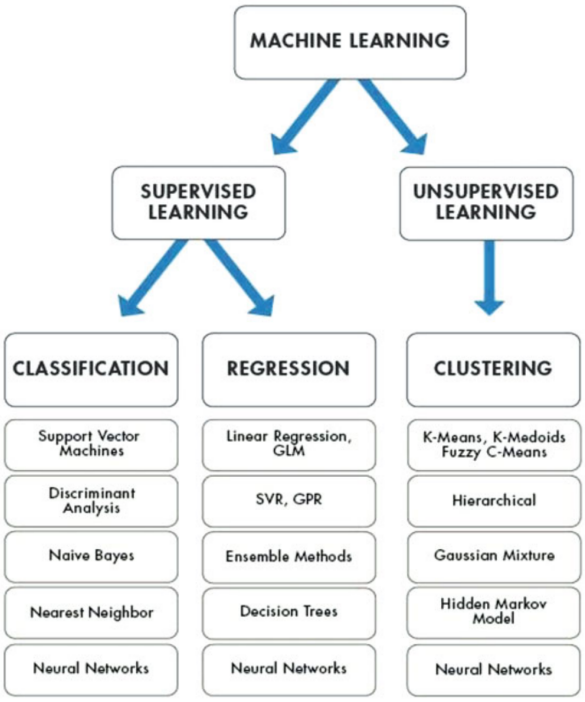
Linear and logistic regression in ML

Admin Córdoba de León

September 2024

1 Linear regression

1.1 Overview



- Supervised learning algorithm.
- One of the most common ways to make inferences and predictions when the target vector is a quantitative value (independent variables can be categorical/discrete).

- Basic predictive analytics technique that uses historical data to predict an output variable.
- Statistical method to establish a relationship between a dependent variable (target vector) and a set of independent variables by fitting the best line/hyper plane.
- Two types of linear regression: simple and multiple.
- β's are known as regression coefficients. They correspond to the **effect of a one-unit change of the independent variables on the target vector** and are assumed to be constant.
- Goal is to find statistically significant values of the parameters β's that minimize the difference between observed and predicted values.
- If we are able to determine the optimum values of these parameters, then we will have the line of best fit that can be used to predict the value of output given a new value of input.
- The best fit line is the one for which the total prediction error is minimum: ordinary least squares (OLS).
- Multiple linear regression can be used to find out which factor has the highest impact on the predicted output and how different variables relate to each other.
- Furthermore, we can use model selection strategies to identify the combination of variables and (their) interactions that produces the best model.
- We can see how well our model performed on the data. The default score for linear regression is R2, which ranges from 0.0 (worst) to 1.0 (best).
- Simple form of neural network having a single layer of learnable parameters.

1.2 Assumptions

Using linear regression requires that the model should conform to the following assumptions:

- The regression model is **linear** in parameters (which are coefficients and the error term).
- Linear regression requires residuals should be **normally distributed**. If the maximum likelihood (not OLS) is used to compute the estimates, then this implies that the dependent and independent variables are also normally distributed.

- The **mean of residuals is zero**. Error term actually refers to the variance present in the response variable that the independent variables failed to explain. The model is said to be unbiased if the mean of the error variable is zero.

- Variance of the residuals is constant, i.e., residuals are evenly distributed around mean or residuals are approximately equal for all predicted dependent variable values. This condition is known as **homoscedasticity**. If the variance changes, it is referred as **heteroscedasticity**.

- Independent variables should not be perfectly correlated (i.e., **no multicollinearity**). Perfect correlation between two variables suggests that they contain same information in them. In other words, both the variables are different forms of the same variable. If variables are correlated, it becomes extremely difficult for the model to determine the true effects of independent variables on dependent variable.

- Residuals should not be correlated with each other. This problem is also known as **autocorrelation**. This is applicable especially for time series data. Autocorrelation is the correlation of a time series with lags of itself. When the residuals are correlated, it means that the current value is dependent on the previous values and that there is a definite unexplained pattern in a dependent variable that shows up in the disturbances.

- Residuals should not be correlated with the independent variables. If residuals are correlated with the independent variable, one can use the independent variables to predict the error. This correlation between error terms and independent variables is known as **endogeneity**. When this kind of correlation occurs, a model may attribute the variance present in error to the independent variable, which in turn produces incorrect estimates.

1.3 Advantages

- Linear regression is one of the most interpretable machine learning algorithms. It is also easy to explain.

- It is easy to use, as it requires minimal tuning.

- It is the most widely used machine learning technique.

1.4 Disadvantages

- Model makes strong assumptions about the data.

- It works on only numeric features, so categorical data requires extra-processing.

- It does not do well with missing values and in the presence of outliers.

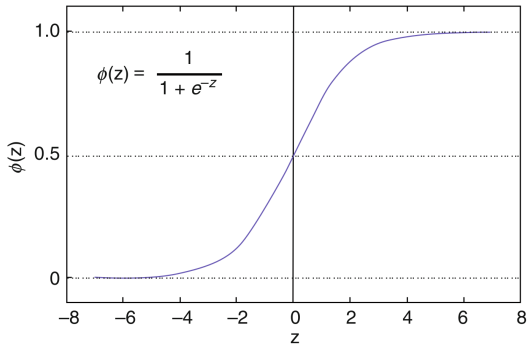
2 Logistic regression

2.1 What is it/How it works?

- Supervised learning algorithm.
- Popular choice for modeling **binary categorical variables**. Logistic regression is the **classification** counterpart of linear regression.
- Predict the probability that an observation is of a certain class. Analyze the effects of a group of independent variables (aka. features) with binary outcomes by **quantifying each independent variable's unique contribution to predict the output of a categorical dependent variable**.
- The outcomes must be categorical or should have discrete values. Usually, the target vector can take only two values (binary classifier). Each can be either yes or no, 0 or 1, true or false, etc.
- Independent variables/features can be either continuous or discrete/categorical.
- Instead of giving the exact value as 0 or 1, it gives the probabilistic values which lie between 0 and 1.

In a logistic regression, a linear model (e.g., $\beta_0 + \beta_1 \mathrm{x}$) of explanatory variables is included in a logistic (also called sigmoid) function,

$$\frac{1}{1 + e^{-}z}$$

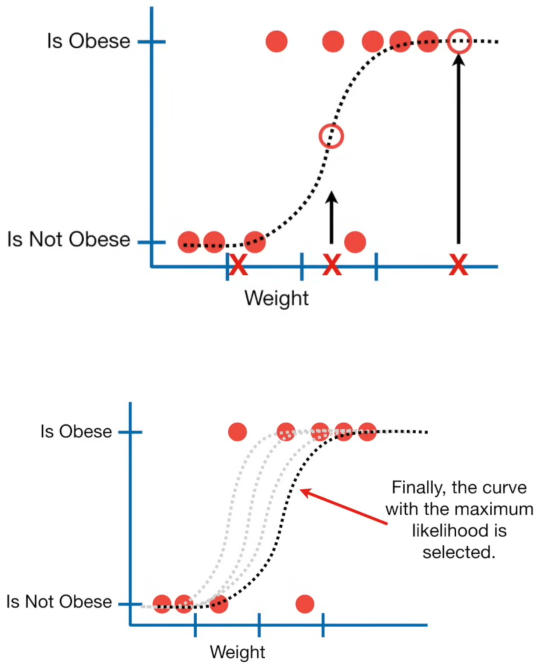


, such that:

$$P(y_i = 1|X) = \frac{1}{1 + e^{-}(\beta_0 + \beta_1 x)}$$

, where $P(y_i = 1|X)$ is the probability of the ith observation's target value, y_i, being class 1; X is the training data; β_0 and β_1 are the parameters to

be learned; and e is Euler's number. The effect of the logistic function is to constrain the value of the function's output to between 0 and 1, so that it can be interpreted as a probability. If $P(y_i = 1|X)$ is greater than 0.5, class 1 is predicted; otherwise, class 0 is predicted. However, the threshold depends on the specific problem we are dealing with.



Since the combinations of explanatory variables are still "linear", the logistic regression works well when classes are linearly separable (i.e., they can be separated by a single decision surface). The goal of the logistic regression is to train the model to find the values of coefficients such that it will minimize the error between the predicted outcome and the actual outcome. These coefficients are estimated using maximum likelihood estimation. In this way, the logistic regression iteratively identifies the strongest linear combination of variables with the highest probability of detecting the observed outcome.

Univariate: $ax + b$

Multivariate: $a_1x_1 + a_2x_2 + ... + a_nx_n + b$

2.2 Types of logistic regression

On their own, logistic regressions are only binary classifiers, meaning they cannot handle target vectors with more than two classes. However, some clever

extensions to logistic regression do just that. In this way, logistic regression can be classified into four types:

- Binomial: In binomial logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, yes or no.
- One-vs-rest (OvR): Heuristic method for using binary classification algorithms for multi-class classification. It involves splitting the multi-class dataset into multiple binary classification problems.
- Multinomial: In multinomial logistic regression, there can be three or more types of dependent (categorical) variables. Multinomial logistic regression can be used to distinguish cats from dogs or sheep.
- Ordinal: In ordinal logistic regression, there can be three or more possible ordered types of dependent variables, such as small, medium, or large.

2.3 Multiclass classifier

In one-vs-rest logistic regression (OvR) a separate model is trained for each class predicted, whether an observation is that class or not (thus making it a binary classification problem). It assumes that each classification problem (e.g., class 0 or not) is independent. Alternatively, in multinomial logistic regression (MLR), the logistic function is replaced with a softmax function:

$$P(y_i = k|X) = \frac{e^(\beta_k x_i)}{\sum_{j=1}^{K} e^(\beta_j x_i)}$$

, where $P(y_i = 1|X)$ is the probability of the ith observation's target value, y_i, being in class k , and K is the total number of classes.

One-vs-rest:

- fit a binary classifier for each class
- predict with all, take largest output
- pro: simple, modular
- con: not directly optimizing accuracy
- common for SVMs as well
- can produce probabilities

"Multinomial" or "softmax":

- fit a single classifier for all classes
- prediction directly outputs best class
- con: more complicated, new code
- pro: tackle the problem directly
- possible for SVMs, but less common
- can produce probabilities

2.4 Reducing variance through Regularization

Regularization is a method of penalizing complex models to reduce their variance. Specifically, a penalty term is added to the loss function we are trying to minimize, typically the L1 and L2 penalties. L2 penalty shrinks all the coefficients towards zero, effectively reducing the impact of each feature. The parameter that determines the strength of regularization is given by C, which takes a default value of 1. Higher values of C correspond to less regularization, in other words, the model will try to fit the data as best as possible. Small values of C correspond to high penalization(or regularization), meaning that the coefficients of the logistic regression will be closer to zero; the model will be less flexible because it will not fit the training data so well. How to find the most appropriate value of C? Usually we need to test different values and see which one gives us the best performance on the test data.

2.5 Assumptions

Logistic regression does not make many of the key assumptions of linear regression and general linear models based on ordinary least squares algorithms. In particular, regarding linearity, normality, homoscedasticity, and measurement level are done away with. First, it does not need a linear relationship between the dependent and independent variables. The primary reason why logistic regression can handle all sorts of relationships is because it applies a nonlinear log transformation to the predicted odds ratio. Second, the independent variables do not need to be multivariate normal, although multivariate normality yields a more stable solution. Also, the error terms (the residuals) do not need to be multivariate normally distributed. Logistic regression does not need variances to be heteroscedastic for each level of the independent variables. Finally, it can handle ordinal and nominal data as independent variables. The independent variables do not need to be metric (interval or ratio scaled). However, some other assumptions still apply which are described below:

- The outcome is a binary or dichotomous variable like true or false, 1 or 0. Reducing an ordinal or even metric variable to dichotomous level loses a lot of information, which makes this test inferior compared to ordinal logistic regression in these cases.
- Logistic regression assumes linearity of independent variables and log odds. While it does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds. In other words, there is a linear relationship between the logit of the outcome and each predictor variable.
- There are no influential values (extreme values or outliers).
- The model should be fitted correctly. Neither over-fitting nor under-fitting should occur. That is only the meaningful variables should be included.

A good approach to ensure this is to use a stepwise method to estimate the logistic regression.

- It requires quite a large sample size. Because maximum likelihood estimates are less powerful than ordinary least squares (e.g., simple linear regression, multiple linear regression). It has been observed that OLS needs 5 cases per independent variable in the analysis, ML needs at least 10 cases per independent variable, and some statisticians recommend at least 30 cases for each parameter to be estimated.
- There is no high intercorrelations (i.e., multicollinearity) among the predictors.

2.6 Model metrics

- It is necessary to encode predicted probabilities as classes in order to use the accuracy score, confusion matrix and ROC curve chart.

2.7 Advantages

- It is easier to inspect and less complex.
- It is a robust algorithm as the independent variables need not have equal variance or normal distribution.
- The algorithm does not assume a linear relationship between the dependent and independent variables and hence can also handle nonlinear effects.
- The algorithm can be regularized to avoid over-fitting.
- The model can be easily updated with new data using stochastic gradient descent.

2.8 Disadvantages

- Logistic regression tends to underperform when there are multiple or nonlinear decision boundaries.
- Model is not flexible enough to naturally capture complex relationships.
- Logistic models tend to over-fit the data when the data is sparse and of high dimension. It requires more data to achieve stability and meaningful results.
- It is not robust to outliers and missing values.

2.9 Ejercicios manuales

1. El siguiente modelo predice la probabilidad de que alguna persona done sangre en una campaña de donación, basados en el género, la edad y la fecha de su última donación. Con base en este modelo, calcula la probabilidad de que una mujer de 72 años de edad que donó sangre hace 120 días, done en la campaña actual. Recuerda que el modelo de regresión logística es una fórmula de regresión lineal "dentro" una función logística.

```
0.545 * gender_F
+ 0.021 * age
-0.001 * time_since_last_gift
-3.39
```

- Female (gender_F=1)
- age 72
- 120 days since last gift

2. Compara la regresión logística con los métodos que hemos revisado hasta el momento, como SVM, Naïve Bayes y Clasificación y clustering.