Berkeley Engineering | BerkeleyHaas

PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Office Hour #13 with Matilde D’Amelio

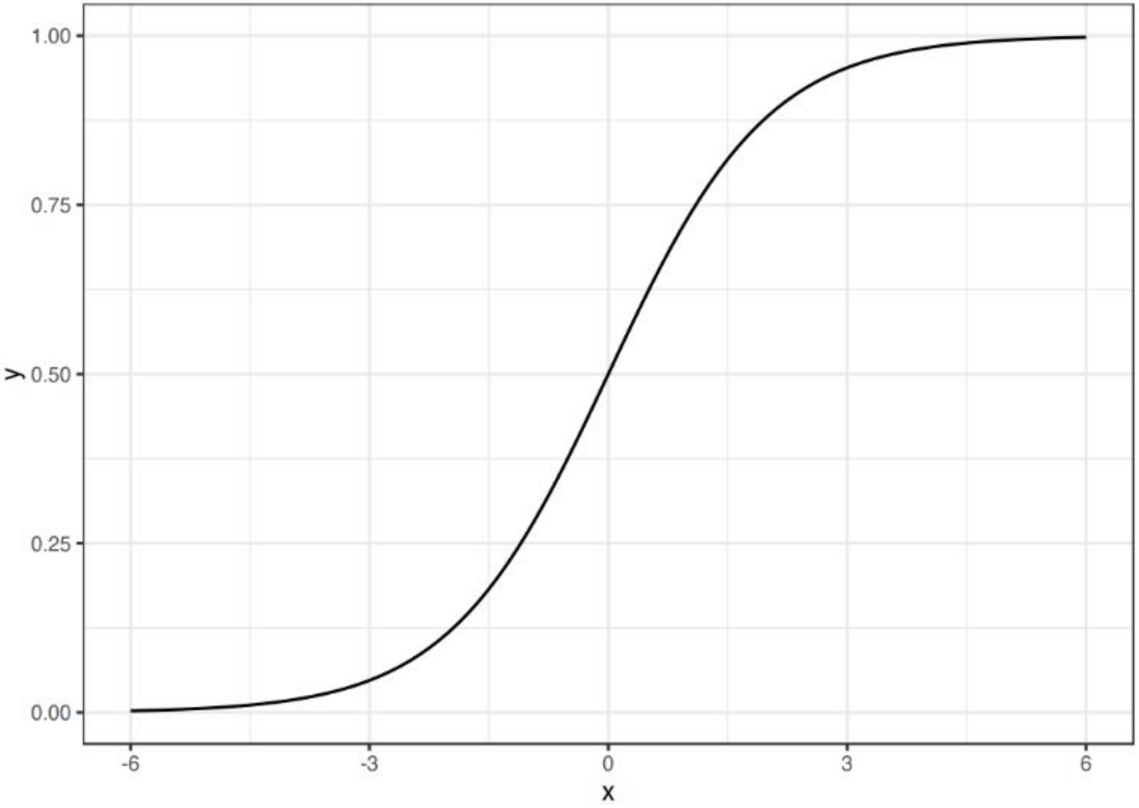
June 16, 2022 at 9 pm UTC

Logistic Regression

It is the go-to method for binary classification problems
For example, if we are modeling people's sex as male or female from their height, then the first class could be male and the logistic regression model could be written as the probability of male given a person's height (Maximum-likelihood estimation**)**

The logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1

$$\text{logistic}(\eta) = \frac{1}{1 + exp(-\eta)}$$



Prepare Data for Logistic Regression

- **Binary Output Variable**: Logistic regression is intended for binary (two-class) classification problems. It will predict the probability of an instance belonging to the default class, which can be snapped into a 0 or 1 classification.
- **Remove Noise**: Logistic regression assumes no error in the output variable (y), consider removing outliers and possibly misclassified instances from your training data.
- **Gaussian Distribution**: Logistic regression is a linear algorithm (with a non-linear transform on output). It does assume a linear relationship between the input variables with the output. Data transforms of your input variables that better expose this linear relationship can result in a more accurate model. For example, you can use log, root, Box-Cox and other univariate transforms to better expose this relationship.
- **Remove Correlated Inputs**: Like linear regression, the model can overfit if you have multiple highly-correlated inputs. Consider calculating the pairwise correlations between all inputs and removing highly correlated inputs.
- **Fail to Converge**: It is possible for the expected likelihood estimation process that learns the coefficients to fail to converge. This can happen if there are many highly correlated inputs in your data or the data is very sparse (e.g. lots of zeros in your input data).

Group Discussion

What can be applications of logistic regressions?

.

Logistic Regression Applications

- **Credit scoring:** It's difficult if you have more than 15 variables in your model. For logistic regression, it is easy to find out which variables affect the final result of the predictions more and which ones less. It is also possible to find the optimal number of features and eliminate redundant variables with methods like recursive feature elimination
- An **e-commerce** company that mails expensive promotional offers to customers, for example, would like to know whether a particular customer is likely to respond to the offers or not: i.e., whether that consumer will be a "responder" or a "non-responder." In marketing, this is called propensity to respond modeling.
- **Healthcare**: discrimination between type 1 and type 2 diabetes in young adult

More details: https://activewizards.com/blog/5-real-world-examples-of-logistic-regression-application

Logistic Regression Many Features

Techniques to Reduce the Number of Features

1. **PCA**: this creates "new" linear combinations of your data where each preceding component explains as much variance in the data as possible. The disadvantage here is that because the components are combinations of your original variables you lose some interpretability with your regression model. It should however produce very good accuracy.
2. **Lasso Regression** uses an L_1 penalization norm that shrinks the coefficients of features effectively eliminating some of them.You can include this L_1 norm into your logistic regression model. It seems Note: Lasso will not explicitly set variable coefficients to zero, but will shrink them allowing you to select the largest coefficients.

Multi-Class Logistic Regression

Eg. If we have to predict whether the weather is sunny, rainy, or windy, we are dealing with a Multi-class problem. We turn this problem into three binary classification problem i.e whether it is sunny or not, whether it is rainy or not and whether it is windy or not. We run all three classifications **independently** on input. The classification for which the value of probability is maximum relative to others, is the solution.

Assumptions:

The Dependent variable should be either nominal or ordinal variable.

Set of one or more Independent variables can be continuous, ordinal or nominal.

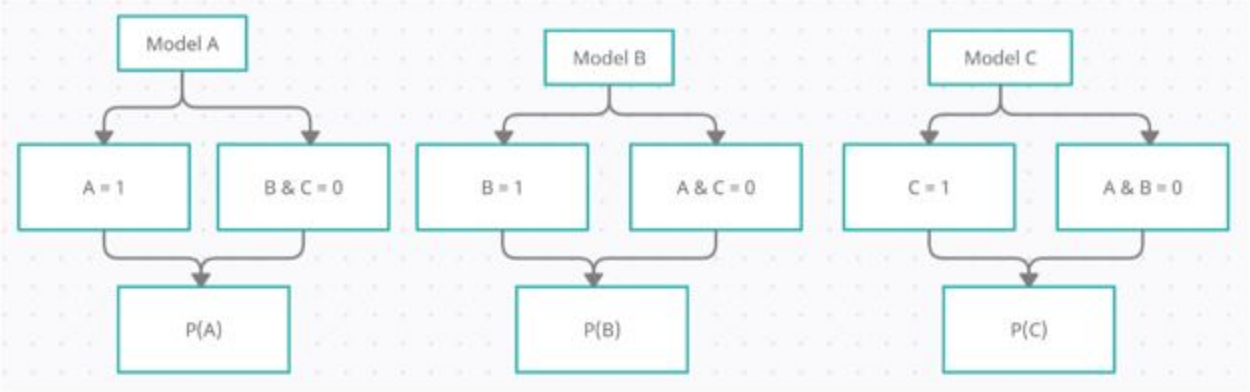
The Observations and dependent variables must be mutually exclusive and exhaustive.

No Multicollinearity between Independent variables.

There should be no Outliers in the data points.

Multi-Class Logistic Regression

K models for K classes



QUESTIONS?