Regression (Application)

1 Overview

Regression in machine learning is a type of predictive modeling technique. It's used for estimating the relationships between a dependent variable (often referred to as the 'outcome' or 'target') and one or more independent variables (known as 'features' or 'predictors'). The main goal of regression is to find a mathematical equation that describes the relationship between the variables.

2 Linear Regression

2.1 Ordinary Least Square Method (OLS)

2.1.1 Objective

The Ordinary Least Squares (OLS) method is a fundamental statistical technique used in linear regression analysis. Its primary objective is to minimize the sum of the squared differences between the observed dependent variable values and those predicted by the linear model. Mathematically, if y_i is the observed value and $\hat{y}_i$ is the value predicted by the model for the i-th observation, the method minimizes the sum $\sum(y_i - \hat{y}_i)^2$.

2.1.2 Linear Model

The OLS method assumes a linear relationship between the dependent variable and one or more independent variables. This relationship is modeled as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \tag{1}$$

2.1.3 Assumptions

1. **Linearity**: The relationship between the independent variables and the dependent variable is assumed to be linear. This means the change in the dependent variable due to a one-unit change in any independent variable is constant

2. **No perfect multicollinearity**: The model should not have perfect multicollinearity, which occurs when one independent variable is a perfect linear combination of others. Multicollinearity can make it difficult to determine the individual effect of each variable and can inflate the variance of the coefficient estimates.

3. **Independence**: Observations are assumed to be independent of each other. In the context of time series data, this means there should be no autocorrelation. If the residuals are correlated, it could suggest that the model is missing some information that is embedded in the structure of the data.

4. **Homoscedasticity**: The residuals (or errors) should have constant variance across all levels of the independent variables. If the variance of the residuals increases or decreases with the independent variables, the model exhibits heteroscedasticity, which violates an OLS assumption.

5. **Exogeneity of Independent Variables**: The independent variables are assumed to be exogenous, meaning they are not correlated with the error term. If this assumption is violated, it can lead to biased and inconsistent parameter estimates.

6. **Normality of error terms**: For the purpose of hypothesis testing and creating confidence intervals, the error terms are assumed to be normally distributed. While OLS itself does not require normality, this assumption allows for the application of various statistical tests after the OLS estimation.

2.1.4 Advantages

1. **Simplicity and Interpretability**: OLS models are easy to understand and interpret, making them a default choice for linear regression problems.

2. **Computational Efficiency**: They are computationally inexpensive, allowing for quick model estimation even with large datasets.

3. **Best Linear Unbiased Estimators**: Provided the assumptions hold, OLS estimators have the lowest variance among all linear estimators.

2.1.5 Limitations

1. **Assumption Heavy**: OLS estimations are reliable only if all underlying assumptions are satisfied, which is often not the case in real-world data.

2. **Outlier Sensitivity:** OLS estimates can be significantly affected by outliers, potentially leading to inaccurate models.

3. **Inability to Model Non-linear Relationships**: Without transformation, OLS cannot capture non-linear relationships between variables.

2.2 Least Absolute Deviation (LAD) Regression

2.2.1 Robust Regression

Robust regression refers to a range of regression techniques designed to be less sensitive to outliers than standard regression methods like OLS. It often down-weights the influence of outliers rather than giving equal weight to all observations as OLS does. Also, instead of minimizing the sum of squared residuals, as in OLS, robust regression

might minimize the sum of absolute residuals or use other loss functions that are less affected by large residuals (e.g., Huber loss or Tukey's biweight loss).

2.2.2 Objective

LAD regression is a robust regression technique. LAD regression minimizes the sum of the absolute values of the residuals. This characteristic makes it less sensitive to outliers than OLS.

2.2.3 Objective Function

The LAD regression solves the following optimization problem:

$$\min \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})| \tag{2}$$

3 Non-linear Regression

3.1 Polynomial Regression

Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an n^{th} degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, and has been used to describe nonlinear phenomena such as the growth rate of tissues, the distribution of carbon isotopes in lake sediments, and the progression of disease epidemics. The math expression is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \varepsilon \tag{3}$$

3.1.1 Non-linear relationship

Despite involving polynomial terms, polynomial regression is still considered a linear model because the regression function is linear in the unknown parameters (coefficients). The non-linearity lies in how the independent variable x is used.

3.1.2 Degree Selection

The degree of the polynomial is a critical parameter in polynomial regression. Too low of a degree, and the model cannot capture the complexity of the data (underfitting). Too high, and the model starts to model random noise in the data (overfitting)

3.1.3 Regularization

To prevent overfitting, techniques such as ridge regression or Lasso regression can be applied, which include a penalty term for large coefficients in higher-degree terms.

4 Gradient Descent

Gradient descent is an optimization algorithm that's used to minimize a function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, gradient descent is typically used to update the parameters of a model in order to minimize the cost function.

4.1 Gradient Calculation

Starting with random values (initial guess) for the model's parameters, gradient descent iteratively adjusts these values in the direction that reduces the cost function.The gradient of the cost function is calculated with respect to each parameter. It gives the direction of the steepest ascent. By moving in the opposite direction, the algorithm seeks the steepest descent.

4.2 Learning Rate

The size of the steps taken during each iteration is determined by the learning rate. A learning rate that's too high can lead to overshooting the minimum, while a learning rate that's too low can slow down the convergence and even get stuck in local minima.

4.3 Different types

4.3.1 Batch Gradient Descent

Computes the gradient of the cost function using the entire dataset. This is computationally expensive and not practical for very large datasets.

4.3.2 Stochastic Gradient Descent (SGD)

Computes the gradient using a single sample at each iteration. This is much faster but can be more erratic in finding the minimum.

4.4 Accelerated Gradient Descent

Accelerated Gradient Descent, also known as momentum gradient descent, is an optimization algorithm designed to **speed up the convergence** of the standard gradient descent algorithm. It achieves this by taking into account the 'momentum' of the gradients, allowing it to navigate the optimization landscape more effectively.

4.4.1 Momentum

In physics, momentum refers to the tendency of an object to continue moving in its current direction. Accelerated Gradient Descent borrows this concept, using the idea of momentum to adjust the parameter updates. Momentum values are typically set between 0 and 1. A common default value for momentum in many neural network libraries is 0.9.

4.4.2 Velocity Update

Along with the gradient, the algorithm maintains a velocity vector, which is a running average of the gradients. This velocity influences the direction and size of the steps taken.

4.4.3 Friction Component

To prevent the velocity from growing too large, a friction component (often termed as 'gamma') is included, which dampens the velocity and ensures stability.