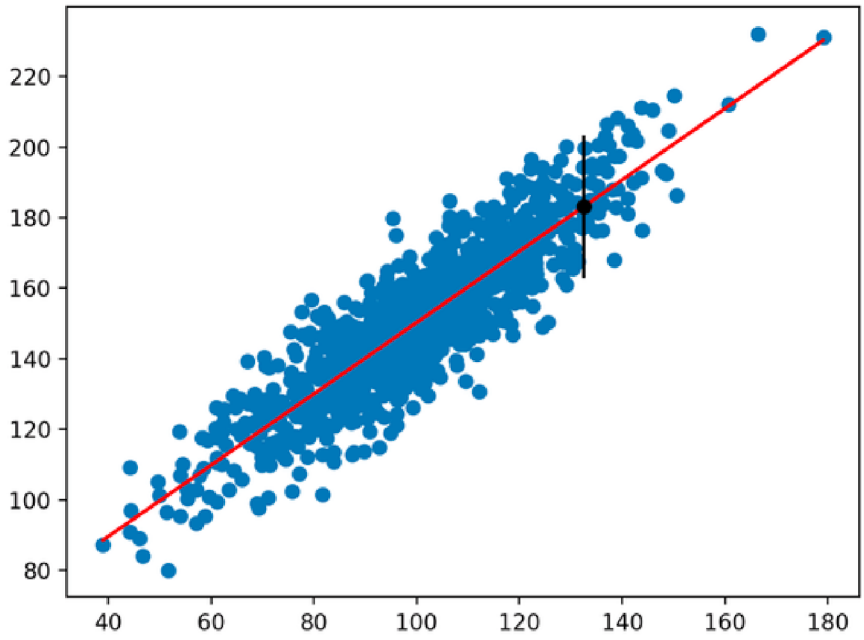
Day 9: Regression Methods



Author: Alisha Metkari

What is Regression

- Regression is supervised machine learning technique used to predict continuous values.
- It uses the relationship between dependent variable and independent variables to predict the outcome.
- The main goal of the regression method is to find the best fit line or curve between dependent variable and independent variables.
- For continuous target variable, there are below basic regression methods. Linear Regression, Multiple Linear Regression, Polynomial Regression.

Author: Alisha Metkari

Best Fit Equation

- The best fit equation is found by minimising the sum of squares of vertical deviations from each data point to the line/curve.
- Least square method is used to find the best fit equation.
- The coefficients of equation are learnt using the gradient descent method.
- The common cost function used is the sum of squared errors.

Author: Alisha Metkari

Linear Regression

- Linear regression finds the linear relationship between dependent and independent variables.
- When only one independent variable is used, it is called Simple Linear Regression.
- The simple linear regression learns the equation of the form, $y = mx + c$, where y is target variable, x is independent variable, m is slope of the line and c is the intercept.
- When more than one independent variables are used, it is called Multiple Linear Regression
- The multiple linear regression learns the equation of the form, $y = a + b*x1 + c*x2 + ...$, where x1 and x2 are the independent variables and a, b, c are coefficients.

Author: Alisha Metkari

Polynomial Regression

- Polynomial regression is the variant of linear regression involving polynomial degree of independent variables.
- It is used to learn the non-linear relationship between data. If there is no linear relationship between dependent and independent variables then polynomial regression is used.
- The original features are transformed to polynomial features of the suitable degree.
- It learns the equation of the form, y = a + b*x + c*x^2. This is the simplest form of the equation.
- In this method, the best fit line is a curve rather than a straight line.
- The degree of polynomial varies from problem to problem and should be chosen accordingly.
- Higher the degree, more the chances of overfitting. Lesser the degree, more the chances of underfitting.

Author: Alisha Metkari

Assumptions of Regression

- For linear regression, the relationship between dependent and independent variables is linear. For polynomial regression, the relationship between dependent and independent variables is curvilinear.
- The mean of the residuals is zero.
- The error terms are uncorrelated with each other.
- The independent variables are uncorrelated with the error term, also called exogeneity
- The error terms have a constant variance, also called homoscedasticity.
- No presence of multicollinearity. No two independent variables are correlated with each other.
- The error terms are normally distributed.

Author: Alisha Metkari

End Notes

- Dependent variable is the target variable, which is the variable that we are predicting.
- Independent variable is the variable on which target variable depends. It is used to predict the target.
- The equation learnt in training is used to predict the unknown dependent variable by providing independent variables.
- Regression models are prone to overfitting.
- Choose the independent variables that are highly correlated with dependent variable and uncorrelated with each other.
- There are more advanced regression methods to handle the drawbacks of the simple regression methods.

Author: Alisha Metkari