Machine Learning (Regression Model)

Prof K R Chowdhary

CSE Dept., MBM University

February 20, 2025

Introduction

⇒ The regression is supervised learning. We consider an example, where we would like to build a model that approximates the relationship f between the number of years of experience in software industry $\mathbf{x}$ and corresponding annual income $\mathbf{y}$.

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_n)$ is (input) years in software industry and $\mathbf{y} = (y_1, y_2, ..., y_n)$ is predicted (output) annual income, f is function describing relationship between $\mathbf{x}$ and $\mathbf{y}$.

⇒ Machine learns f given $\mathbf{x}$, $\mathbf{y}$. The ε is *random error term* either positive or negative with mean zero, and represents irreducible error in the model, which is the theoretical limit around the performance of the algorithm due to inherent noise.

⇒ Two tasks of supervised learning are: 1) Regression, which predicts a continuous numerical value, and 2) to assign a label, for example, to predict whether a given picture is of a "cat" or "dog?"

Regression Model

⇒ In regression, data is split into *training data* set and *test data* set. Goal is: to learn linear model using *ordinary least squares regression* that predicts a new y given a previously unseen x, with as little error as possible.

⇒ It is a parametric method, to find a function that predicts $\hat{y}$ for given specific x:

$$\hat{y} = \beta_0 + \beta_1 * x + \varepsilon, \tag{2}$$

here β_0 is y-intercept (point where line cuts y axis), $\beta_1 =$ slope of line, i.e., how much it increases or decreases by one year of experience.

⇒ Goal: learn model parameters β_0 and β_1 that minimizes *error*.

⇒ To find best value of parameters:

1. Define a *cost function* (or "loss function"), that measures how inaccurate the predictions are?
2. Find the parameters that minimize cost, i.e., make this as accurate as possible.

Regression Model

⇒ In 2D it is a line of best fit, in 3D it is plane, ...

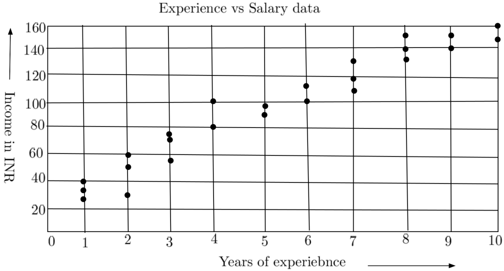
⇒ Mathematically, we look at difference between each real data point (y) and this model's prediction ($\hat{y}$).

⇒ Differences are squared to avoid negative numbers, and we penalize large differences. At end, all squares are summed up and averaged: a measure of how well our data fits a line. For n number of observations:

$$Cost = \frac{1}{2n}\sum_{1}^{n}((\beta_1 x_i + \beta_0) - y_i))^2 \tag{3}$$

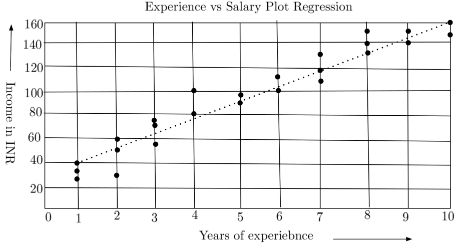
"Cost" should be minimum possible. Using $2 * n$, instead of n makes the mathematics workout more cleanly when taking derivative to minimize loss (page no. 7). The random error term ε is not accounted for in equation (3), its mean value is zero.

⇒ *Example.* Let an attribute vector $\mathbf{x} = [1, 1, 1, 2, ..., 10, 10]$ is years of experience of software developers, and vector $\mathbf{y} = [26, 33, 40, 30, ..., 144, 160]$ is corresponding annual income in thousands of INR. The values are plotted in Fig-1 below, as "Experience vs Salary."



⇒ Linear regression curve is obtained in Fig. (next slide) for experience vs salary data shown in Fig. 1, using cost equ. (3). The β_0 and β_1 are chosen such that *cost* is minimum, for all input data of experience vs salary.

Fig.2 (Experience vs Salary Plot Regression: Manually created).



⇒ Values of β_0, β_1 are substituted in equation (2) to get linear relation, which is *induced classifier*, and can classify new data in similar way we did in liner classifier in equation (4).

$$2.5 - 0.8x_1 - x_2 = 0 \tag{4}$$

For this simple problem we can compute a closed form solution using calculus to find optimal β parameters that minimize cost function (eq. 3). As cost function grows in complexity, finding a closed form calculus is difficult (requires different method).

Calculating β_1 and β_0

- Method: Ordinary Least Squares (OLS) Estimation finds the coefficients of a linear regression model by minimizing sum of squared differences (residuals) between observed y_i values and predicted $\hat{y}$ values.
- OLS is closed-form solution of the linear regression problem, derived by taking derivative of the sum of squared residuals with respect to coefficients (β_0 and β_0) and setting it equal to zero.

The formulas are:

$$\beta_1 = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \tag{5}$$

$$\beta_0 = \frac{\sum y_i - \beta_1 \sum x_i}{n} \tag{6}$$

- These equations give a direct way to compute the slope and intercept for the best-fitting line (attached program linreg.py), and are typically used when solving problem without need for iterative techniques (like gradient descent).

Plotting regression line

The computed regression line (y_i for all x_i) is: x = [1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9, 9, 10, 10, 10], y = [26, 33, 40, 40, 47, 54, 56, 63, 70, 72, 79, 86, 88, 95, 102, 104, 111, 118, 120, 127, 134, 141, 144, 151, 158, 157, 160, 160, 163, 170].

(7)

Applications of Regression

1. **Predictive Modeling**

Forecasting Sales: Companies use regression to predict future sales based on historical data.

Stock Market Prediction: Regression models can help predict stock prices by analyzing historical data and identifying patterns and trends,

Weather Forecasting: Meteorologists use regression to predict weather patterns,

2. **Economics and Business**

Demand Forecasting: Economists and businesses use regression to predict demand for products based on factors such as price, income levels,

Cost and Profit Analysis: Businesses use regression to understand the relationship between various cost factors (e.g., production costs, labor costs) and profit.

Risk Assessment: Banks and financial institutions use regression to assess the risk associated with loan applications

3. Healthcare and Medicine

Disease Prediction and Diagnosis: To predict the likelihood of a disease occurring based on various factors such as age, gender, ...,

Treatment Effectiveness: To analyze how effective different treatments are,

*Epidemiological Studies:*To study the relationship between environmental factors, lifestyle, ...

4. Marketing and Advertising

Customer Behavior Analysis: To understand how various factors (e.g., price, advertisement, product features)

5. Education

Student Performance Prediction: To predict students' academic performance based on factors like attendance, study hours, family background, and previous grades.

Evaluating the Impact of Teaching Methods: Can assess how different teaching methods, curricula, or tools affect student performance.

Dropout Prediction: To predict the likelihood of students dropping out based on ...

6. Engineering and Manufacturing

Quality Control: To determine the relationship between production variables (e.g., temperature, material type) and product quality, to optimize production process.

Predictive Maintenance: Can predict when a machine or system will require maintenance, based on variables like operating hours, wear, ...

7. Environmental Studies

Climate Change Modeling: Researchers use regression to model the relationship between human activities (e.g., carbon emissions) and climate change variables like temperature or sea levels.

Pollution Impact Analysis: Regression can be used to understand the impact of pollution on health, the environment, or biodiversity by analyzing data on emissions, population exposure, and disease rates.

8. Agriculture

Crop Yield Prediction: Regression is used in agriculture to predict crop yields based on factors such as soil quality, weather patterns, and irrigation techniques.

Farming Optimization: Regression can help farmers optimize planting schedules, fertilizer use, and irrigation to maximize crop production and reduce costs.