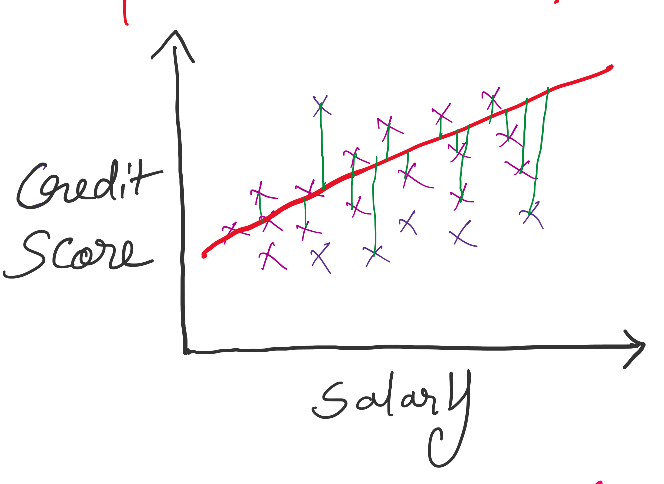
Linear Regression

Saturday, 8 March 2025 8:29 PM

→ Given a set of n datapoints in d dimensions, the task is to fit a straight line passing through the points. The line should fit the data in the "best possible manner".



→ In the ML setup typically you are given a matrix $A \in \mathbb{R}^{n \times d}$.
Each row of the matrix represents a data point with d features/attributes.
Along with A, we are also given a vector $b \in \mathbb{R}^n$ containing the labels.
The task is to find a vector $x \in \mathbb{R}^d$ s.t. the quantity $\|Ax - b\|_2^2$ is minimized.

Linear Regression:

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2$$

Example:- House Price Prediction

$$\begin{array}{c} \\ \text{House 1} \\ \text{House 2} \\ \text{House 3} \\ \vdots \\ \text{House } n \end{array}
\underbrace{\left[\begin{array}{c} \text{No. of Rooms} \quad \text{Carpet Area} \quad \text{Distance to Nearest Hospital} \quad \text{Distance to School} \\ \text{———} \; a_1^T \; \text{———} \\ \text{———} \; a_2^T \; \text{———} \\ \\ \\ \\ \end{array}\right]}_{A}
\approx
\begin{array}{c} \text{Price of House in Lakhs} \\ \underbrace{\begin{bmatrix} b_1 = 25 \\ b_2 = 48 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}}_{b} \end{array}$$

→ Suppose $n = 1000$, $d = 4$
So we have the rows representing 1000 houses each having 4 attributes, b represents the prices of 1000 houses

→ We assume that the price of the house is a linear function of its features
i.e. for house a_1

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 \approx b_1$$

So, we want to find $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$ s.t. $Ax \approx b$

Lets say $Ax = b'$

Then error is given by

$$(b_1' - b_1)^2 + (b_2' - b_2)^2 + \dots (b_n' - b_n)^2$$

i.e.

$$\text{loss} = \sum_{i=1}^{n} (b_i' - b_i)^2 = \sum_{i=1}^{n} (a_i^T x - b_i)^2 = \|Ax - b\|_2^2$$

Hence in Linear Regression, we try to minimize the above loss so the problem becomes

$$\min_{x \in \mathbb{R}^d} \|Ax - b\|_2^2$$

or

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{n} (a_i^T x - b_i)^2$$

Here let

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^d} \|Ax - b\|_2^2$$

then x^* is our ML Model.

How to predict the price of a new house a_{n+1}^T?

How to get x^*?
→ There is a closed form solution
→ Gradient Descent

→ Data Normalization Required in practice
Feature Scaling

$$\tilde{a}^k = \frac{a^k - \mu}{\sigma} \quad \text{(Z-Score normalization)}$$