CS 217: Artificial Intelligence and Machine Learning **Jan-Apr 2024**

Lecture 3: Regression

Lecturer: Swaprava Nath *Scribe(s): SG5, SG6*

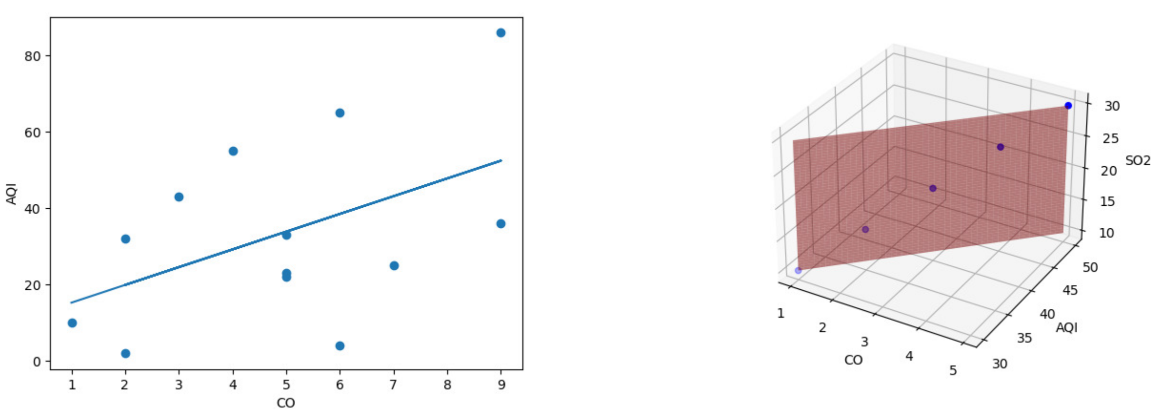
Disclaimer: *These notes aggregate content from several texts and have not been subjected to the usual scrutiny deserved by formal publications. If you find errors, please bring to the notice of the Instructor.*

3.1 Introduction

Regression analysis finds a practical application in everyday life through the *Air Quality Index (AQI)*, a tool designed to quantify air pollution. Given the challenges of directly measuring and calculating the percentages of each air component like SO_2 and CO, regression comes into play
Instead of individually measuring all components, a subset is measured, and the remaining components are estimated using regression analysis. This statistical technique enables the establishment of relationships between measured and unmeasured components, offering a more efficient means of interpreting air quality by inferring the percentages of various pollutants without the need for exhaustive measurements.

$$AQI = \max\{f_1(x_1), f_2(x_2), \dots f_n(x_n)\}$$

Where $f_i(x_i)$ is an unique function for each pollutant



(a) AQI vs CO (b) AQI vs SO2 and CO

Figure 3.1: function of various pollutant concentration

In this case device might find only 1 or 2 functions and based on that guess the AQI i.e. to fit the perfect AQI data with limited observations and estimate AQI value.

3.2 Linear Regression

We use *Linear Regression* for estimating an unknown data from a know data as

1. It is a simple and powerful tool
2. It is interpretable
3. It's works on transformations of raw data

How do we best fit the given data?
We need a measurement criteria to calculate goodness of our estimation function. So we use an *error function* also called as *loss function, lost function, energy function.* It has two parameters *estimation function* and *data points*

$$\text{Error function} = E(f, D)$$

$$f \textit{ is the estimation function}$$

$$D = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\} \text{ where } (x_i, y_i) \text{ is a data point from the data}$$

3.2.1 Possible Error functions

$$\sum_{i=1}^{n}(f(x_i) - y_i) \tag{3.1}$$

- 3.1 is not a good error function as it is signed.

$$\sum_{i=1}^{n}|f(x_i) - y_i| \tag{3.2}$$

- 3.2 is a better error function than 3.1 as it is unsigned.

$$\sum_{i=1}^{n}(f(x_i) - y_i)^2 \tag{3.3}$$

- 3.3 is squared cost function, most used error function

$$\sum_{i=1}^{n}(f(x_i) - y_i)^3 \tag{3.4}$$

- 3.4 is not a good error function as it is signed

3.2.2 Squared loss function

$$\sum_{i=1}^{n}(f(x_i) - y_i)^2$$

1. It is a *continuous function* and in particular *differentiable*.
2. Easy to visualize in Euclidean space.
3. Mathematical analysis become easier.

Let DS be the data set

$$DS = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$$

Where x_i is the input and y_i is the output for the i^{th} training example. The number n = number of data samples or more usually called training instances. $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Here $\mathbb{R}^d$ is the d-dimensional space.

Let PM 2.5, SO_2, CO be the components of x_i then,

$$x_i = \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ x_{i_3} \end{bmatrix}$$

x_{i_1} represents the concentration of PM 2.5
x_{i_2} represents the concentration of CO
x_{i_3} represents the concentration of SO_2

Let us define a X matrix containing x_i

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1d} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2d} \\ x_{31} & x_{32} & x_{33} & \ldots & x_{3d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \ldots & x_{nd} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ x_3^T \\ \vdots \\ x_n^T \end{bmatrix}_{nxd}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{nx1}$$

3.2.3 General Regression

The goal of this is to find a function $f^*(x)$ such that it is the first prediction of y (output data) w.r.t. D.

$$f^* \in \arg\min E(f, D)$$
subject to $f \in \mathcal{F}$, where $\mathcal{F}$ is the set of all functions

3.2.4 Parameterized Regression

In this f is a function of the form $f(x, w)$, where w are the parameters of regression.
e.g. $f(x,(\alpha,\lambda)) = \alpha e^{\lambda^T x}$
e.g. $f(x, w) = \sum w_i x_i$ i.e., $f(x, w) = w_0 + w_1 x + w_2 x^2 + ... + w_k x^k$

We can use parameterized regression to reduce the solution space we needed to search from in case of a general regression. Let us take an example of parameterized function:

- $f(\underline{x},(\alpha,\underline{\lambda})) = \alpha e^{-\lambda^T x}$

In parameterized regression we need to minimize the parameterized function w.r.t the given parameters i.e.,

$$f \equiv f(x, w)$$
$$\arg\min_w E(f(x, w), D)$$

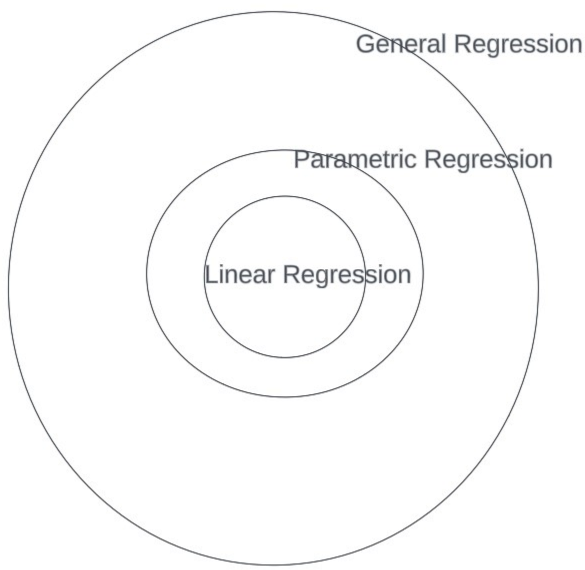


Figure 3.2: Diagram Representation

3.2.5 Linear Regression

In this f is a function of the form $f(x, w) = w^T x + w_0 = \bar{w}^T x$ here $w \in \mathbb{R}^d$

3.3 Least Square Optimisation for Linear Regression

$$W^* \in \arg\min_w \left(\sum_{i=1}^{n} \left(\sum_{j=0}^{d} w_j x_{ij} - y_i \right)^2 \right)$$

For $d = 1$

$$E(w, D) = \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2$$

$$\begin{aligned}
\frac{\partial E}{\partial w_0} &\implies -2 \left(\sum_{i=1}^{n} (y_i - w_0 - w_1 x_i) = 0 \right) \\
&\implies \sum y_i - n w_0 - w \sum x_i = 0 \\
\frac{\partial E}{\partial w_1} &\implies -2 \left(\sum_{i=1}^{n} x_i (y_i - w_0 - w_1 x_i) = 0 \right) \\
&\implies \sum x_i y_i - w_0 \sum x_i - w_1 \sum x_i^2 = 0
\end{aligned}$$

Note that the 2 equations above are a linear equations in variables w_0 and w_1. Solving for these we get,

$$\begin{aligned}
w_1 =& \frac{n * \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \\
w_0 =& \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}
\end{aligned}$$

3.3.1 Case 1: d=1

$$E(w, d) = \sum_{i=1}^{n} \left(y_i - w_0 - w_1 x_i \right)^2$$

Find w_0 w_1 such that

$$\frac{\partial E}{\partial w_0} = 0 \tag{3.5}$$

$$\frac{\partial E}{\partial w_1} = 0 \tag{3.6}$$

From equation 3.5

$$w_0 = \frac{\sum y_i - w_1 \sum x_i}{n} \tag{3.7}$$

From equation 3.6

$$w_1 = \frac{\sum x_i y_i - w_0 \sum x_i}{\sum {x_i}^2} \tag{3.8}$$

Take $\alpha = \frac{\sum x_i y_i}{\sum {x_i}^2}$ and $\beta = \frac{\sum {x_i}^2}{n}$

let $\bar{x}$ be $\frac{\sum x_i}{n}$ and $\bar{y}$ be $\frac{\sum y_i}{n}$

$$w_1 = \alpha - w_0 * \frac{\bar{x} * n}{n * \beta} \tag{3.9}$$

$$w_1 = \alpha - (\bar{y} - w_1 \bar{x}) \frac{\bar{x} * n}{n\beta} \qquad From \quad 3.7 \tag{3.10}$$

$$w_1 * (1 - \frac{\bar{x}^2}{\beta}) = \alpha - \frac{\bar{y}\bar{x}}{\beta} \tag{3.11}$$

$$w_1 = \frac{\alpha * \beta - \bar{x} * \bar{y}}{\beta - \bar{x}^2} \tag{3.12}$$

Exercise: Find $\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ in terms of α and β (It turns out to be w_1)

3.3.2 Case 2: For d-dimensional data

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix} \qquad X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Let $z_i = y_i - w^T x_i$

$$w^* \in \arg\min_w \sum_{i=1}^{n} (y_i - w^T x_i)^2 = \sum_{i=n}^{n} z_i^2 = \|z\|^2$$

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} y_1 - x_1^T w \\ y_2 - x_2^T w \\ \vdots \\ y_n - x_n^T w \end{bmatrix} = y - Xw$$

$$\|z\|^2 = \|y - Xw\|^2$$

$$w^* \in \arg\min_w \|y - Xw\|^2$$

$$\|y - Xw\|^2 = (Xw - y)^T (Xw - y)$$

$$\arg\min_w (Xw - y)^T (Xw - y) = (w^T X^T - y^T)(Xw - y)$$

$$E(w, D) = w^T X^T X w - w^T X^T y - y^T X w + y^T y$$

$$(w^T X^T y = y^T X w)$$

$$E(w, D) = w^T X^T X w - 2y^T X w + y^T y$$

We can find w by doing $\nabla_w E = 0$

So $\frac{\partial(2y^T X w)}{\partial w} = (2y^T X)^T = 2X^T y$
$\frac{\partial(w^T X^T X w)}{\partial w} = X^T X w + (X^T X)^T w = X^T X w + X^T X w = 2X^T X w$
$\frac{\partial(y^T y)}{w} = 0$

$$\nabla_w E = 0 \implies 2X^T X w - 2X^T y = 0$$
$$\implies 2X^T X w = 2X^T y$$
$$\implies \boxed{w = (X^T X)^{-1} X^T y}$$

$$\boxed{(w^*)^T x = \hat{y}}$$

If $X^T X$ is not invertible then it means that the closed form expression cannot be used to find the optimal w^*. A possible such scenario is when there are less data points than the dimension of the data.