Regression

Instructor: Prof. Ganesh Ramakrishnan

Recap

- Supervised (Classification and Regression) vs Unsupervised Learning
  - Three canonical learning problems
- What is data and how to predict
  - More on this today in the context of regression
- Squared Error

Agenda

- What is Regression
- Formal Defintion
- Types of Regression
- Least Square Solution
- Geometric Interpretation of least square solution

Regression

- Finding correlation between a set of output variables and a set of input variables (x)

  so far single output variable (y)
- Input variables are called *independent variables*
- Output variables are called *dependent variables*

$y = f(x)$ ... Linear regression, f is linear

Examples

- A company wants to how much money they need to spend on T.V advertising to increase sales to a desired level, say y*
- They have previous data of form $<x_i, y_i>$, where x_i is money spent on advertising and y_i are sale figures
- They now fit the data with a function, lets say linear function

$\therefore$ To get Δy increase in y, you need $\Delta y / \beta_1$ increase in x ←

$$y = \beta_0 + \beta_1 * x \quad \text{so that} \quad y_i \approx \beta_0 + \beta_1 x_i \ \forall i \tag{1}$$

and then find the money they need to spend using this function

- Regression problem is to find the appropriate function and its coefficients

We will replace x by $\phi(x)$

one-many

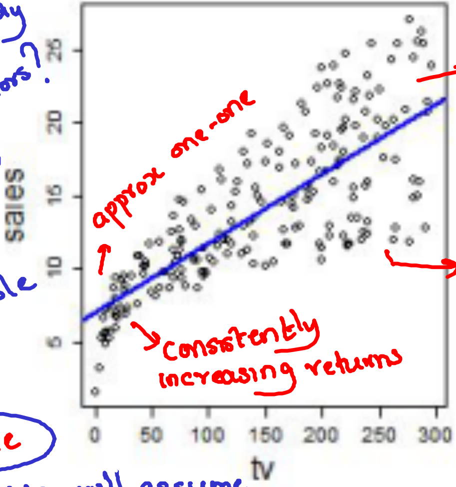
probably other factors?

0000 → many-one

This factor is not playing a significant role in this region

0000 → one-one

We will assume this



approx one-one

consistently increasing returns

many-many

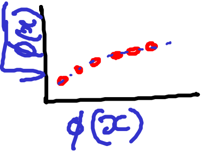
somewhat inconsistent

Figure: Linear regression on T.V advertising vs sales figure

What if sales is a non-linear function of advertising?

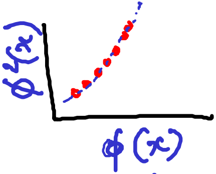
: $\underbrace{Y}_{\text{sales}} = \beta_0 + \beta_1 \underbrace{\phi(x)}_{\text{investment in advertisement.}}$

x = date for TV company



ie Y varies approximately as $\sqrt{\phi(x)}$

$$Y = \beta_0 + \beta_1 \sqrt{\phi(x)}$$



OR Y varies approximately as $\phi^2(x)$

$$Y = \beta_0 + \beta_1 \phi^2(x)$$

In reality, could be a combination!

$$Y = \beta_0 + \beta_1 \phi(x) + \beta_2 \sqrt{\phi(x)} + \beta_3 \phi^2(x) + \cdots$$

we want to learn $\beta_0, \beta_1, \beta_2 \ldots$

We assume 1-1 mapping between Y & x (approx mapping)

Given n observations:

$$\begin{bmatrix} y_1, & \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_p(x_1) \\ y_2, & \phi_1(x_2), & \phi_2(x_2) & \cdots & \phi_p(x_2) \\ \vdots & & & & \\ y_n, & \phi_1(x_m), & \phi_2(x_m) & \cdots & \phi_p(x_m) \end{bmatrix}$$

x_i = corresponds to day i for the TV company

$\phi_1(x_i), \phi_2(x_i) \cdots \phi_p(x_i)$ are different attributes of the T.V company on day i

y_i is the amount of sales on day i

We need to estimate $w_0, w_1, w_2 \ldots w_k$ ($\beta_1, \beta_2 \ldots$ on previous slide) such that

$$y_i \approx w_0 + w_1 \underbrace{\phi_1(x_i)}_{\phi_1(x_i) = ad} + w_2 \underbrace{\phi_2(x_i)}_{\sqrt{\phi_1(x_i)}} + \cdots + w_k \underbrace{\phi_k(x_i)}_{\phi_1^2(x_i)}$$

For eg, on prev slide: $\phi_1(x_i) = ad$ investment on ith day $\sqrt{\phi_1(x_i)}$ $\phi_1^2(x_i)$

Some more notes:

1. $\phi_1 \ldots \phi_K$ can be any attribute that has the potential of linearly influencing y

2. $\phi_1(x_i), \phi_2(x_i) \ldots \phi_p(x_i)$ may or may not be functions of each other...
   But certain machine learning approaches are robust to interdependencies among ϕ_j's and certain others are NOT

3. You could use ML workbench such as WEKA (Waikato univ, New Zealand) to study ϕ_j's that "linearly influence" y.

④ How to deal with the attribute:

Celebrity in ad is → Actor, Actress, Academician, Student, Common person

Example of nominal attribute

$$\left.\begin{array}{l}\phi_{c=ar}(x_i)\\ \phi_{c=as}(x_i)\\ \phi_{c=cp}(x_i)\\ \phi_{c=s}(x_i)\end{array}\right\}$$

For a given i exactly one of these can take value $=1$ and the rest take value $=0$

Nominal attribute: One that takes one value from a set of discrete possible values.

Formal Definition

- Two sets of variables: $x \in \mathcal{R}^N$ (independent) and $y \in \mathcal{R}^k$ (dependent)
- D is a set of m data points: $<x_1, y_1>, <x_2, y_2>, ..., <x_m, y_m>$
- ϵ (f, D): An error function, designed to reflect the discrepancy between the predicted value $f(x_i)$ and $y_i \ \forall i$ $f(x_i) = \omega_0 + \omega_1 \phi_1(x_i) + \omega_2 \phi_2(x_i) \ldots$
- Regression problem: Determine a function f^* such that $f^*(x)$ is the best predictor for y, with respect to D,

$$f^* = \underset{f \in F}{\operatorname{argmin}} \ \epsilon(f, D) \tag{2}$$

where, F denotes the class of functions over which the optimization is performed

Types of Regression

- Depends on the function class and error function
- Linear Regression : establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line, i.e

$$Y = a + b * X \tag{3}$$

  - Here F is of the form $\Sigma_{i=1}^{p} w_i \phi_i(x)$, where ϕ_i are called basis functions (or attributes)
  - Problem is to find w^* where

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmin}}\ \epsilon(\mathbf{w}, \mathbf{D}) \tag{4}$$

- Ridge Regression : A shrinkage parameter (regularization parameter) is added in the error function to reduce discrepancies due to variance → Linear regression with good generalization
- Logistic Regression : Used to model conditional probability of dependent variable given independent variable and is extensively used in classification tasks

$$log\frac{p(y|x)}{1-p(y|x)} = \beta_0 + \beta * x \quad (5)$$

→ Different function class

- Lasso regression, Stepwise regression and many more

Different error function

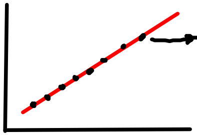
Least Square Solution

- Form of ϵ plays a major role in the accuracy and tractability of the optimization problem
- The squared loss is a commonly used error/loss function. It is the sum of squares of the differences between the actual value and the predicted value

$$\epsilon(f, D) = \sum_{j=1}^{m} (f(x_j) - y_j)^2 \tag{6}$$

- The least square solution for linear regression is given by

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{j=1}^{m} \left(\Big(\underbrace{\sum_{i=1}^{p} w_i \phi_i(x_j)}_{f(x_j)} - y_j \Big)^2 \right) \tag{7}$$



$$f(x_j) = \underbrace{w_1^* \phi_1(x_j) + \dots w_p^* \phi_p(x_j)} = y_j$$

assume that some $\phi_k(x_j) = 1$
to account for offset/bias
of linear function

- The minimum value of the squared loss is zero
- If zero were attained at $\mathbf{w}^*$, we would have ....

Suppose

$$\sum_{i=1}^{p} w_i^* \phi_i(x_j) = y_j \quad \forall j = 1 \dots m$$

- The minimum value of the squared loss is zero
- If zero were attained at $\mathbf{w}^*$, we would have $\forall u, \phi^T(x_u)\mathbf{w}^* = \mathbf{y_u}$, or equivalently $\underline{\phi\mathbf{w}^* = \mathbf{y}}$, where

$$\underline{\phi} = \begin{bmatrix} \phi_1(x_1) & \dots & \phi_p(x_1) \\ \dots & \dots & \dots \\ \phi_1(x_m) & \dots & \phi_p(x_m) \end{bmatrix}$$

each row is for an example

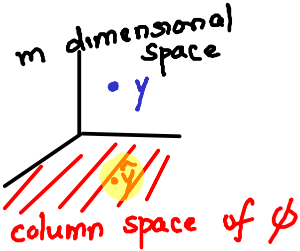
and

$$\underline{\mathbf{y}} = \begin{bmatrix} y_1 \\ \dots \\ y_m \end{bmatrix}$$

- It has a solution if $\mathbf{y}$ is in the column space (the subspace of R^n formed by the column vectors) of ϕ : Obtain w^* using Gaussian elimination

- The minimum value of the squared loss is zero
- If zero were NOT attainable at $\mathbf{w}^*$, what can be done?

Least squares:
- Consider all $\hat{y}$'s in column space of ϕ matrix
- Find that $\hat{y}$ which is as close as possible to observed y
- Thus, $\hat{y}$ in column space st $(y-\hat{y})$ is $\perp$ to column space is the LS soln



Geometric Interpretation of Least Square Solution

- Let $\mathbf{y}^*$ be a solution in the column space of ϕ
- The least squares solution is such that the distance between $\mathbf{y}^*$ and $\mathbf{y}$ is minimized
- Therefore........... $y^* = \phi w$

$(y - y^*)^T \phi = 0$

$a \perp b$ iff $a^T b = 0$

Geometric Interpretation of Least Square Solution

- Let $\mathbf{y}^*$ be a solution in the column space of ϕ
- The least squares solution is such that the distance between $\mathbf{y}^*$ and $\mathbf{y}$ is minimized
- Therefore, the line joining $\mathbf{y}^*$ to $\mathbf{y}$ should be orthogonal to the column space

$$\phi\mathbf{w} = \mathbf{y}^* \tag{8}$$

$$\phi^T(\phi w) = \phi^T y^* = \phi^T y$$

$$(\mathbf{y} - \mathbf{y}^*)^{\mathbf{T}}\phi = \mathbf{0} \tag{9}$$

$$(\mathbf{y}^*)^{\mathbf{T}}\phi = (\mathbf{y})^{\mathbf{T}}\phi \tag{10}$$

$$(\phi^T\phi)\,w = \phi^T y \Rightarrow w = (\phi^T\phi)^{-1}\phi^T y$$

$$(\phi \mathbf{w})^{\mathbf{T}} \phi = \mathbf{y}^{\mathbf{T}} \phi \tag{11}$$

$$\mathbf{w}^{\mathbf{T}} \phi^{\mathbf{T}} \phi = \mathbf{y}^{\mathbf{T}} \phi \tag{12}$$

$$\phi^{T} \phi \mathbf{w} = \phi^{\mathbf{T}} \mathbf{y} \tag{13}$$

$$\mathbf{w} = (\phi^{\mathbf{T}} \phi)^{-\mathbf{1}} \mathbf{y} \tag{14}$$

- Here $\phi^T \phi$ is invertible only if ϕ has full column rank

Proof?

Theorem : $\phi^T\phi$ is invertible if and only if ϕ is full column rank
Proof :
Given that ϕ has full column rank and hence columns are linearly independent, we have that $\phi\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{x} = \mathbf{0}$
Assume on the contrary that $\phi^T\phi$ is non invertible. Then $\exists\mathbf{x} \neq \mathbf{0}$ such that $\phi^T\phi\mathbf{x} = \mathbf{0}$

$$\Rightarrow \mathbf{x^T}\phi^\mathbf{T}\phi\mathbf{x} = \mathbf{0}$$
$$\Rightarrow (\phi\mathbf{x})^\mathbf{T}\phi\mathbf{x} = \mathbf{0}$$
$$\Rightarrow \phi\mathbf{x} = \mathbf{0}$$

This is a contradiction. Hence $\phi^T\phi$ is invertible if ϕ is full column rank
If $\phi^T\phi$ is invertible then $\phi\mathbf{x} = \mathbf{0}$ implies $(\phi^T\phi\mathbf{x}) = \mathbf{0}$, which in turn implies **x = 0 , This implies ϕ has full column rank if $\phi^T\phi$ is invertible. Hence, theorem proved**

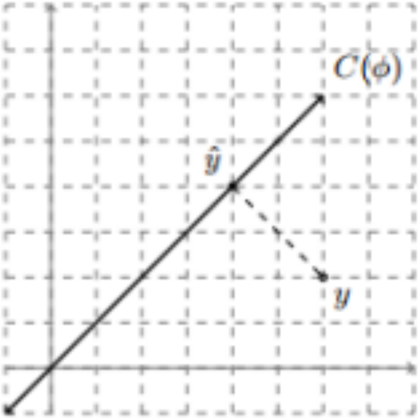


Figure: Least square solution $\mathbf{y}^*$ is the orthogonal projection of y onto column space of ϕ