Lecture 3 - Regression

Instructor: Prof. Ganesh Ramakrishnan

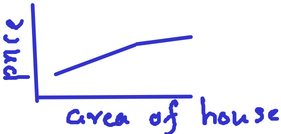
The Simplest ML Problem: Least Square Regression

- Curve Fitting: Motivation
  - Error measurement
  - Minimizing Error
- Method of Least Squares

Simplest is least squares, bayesian linear reg, regularized error, parametric vs non-param

linear algebra, optimization

Curve Fitting: Motivation

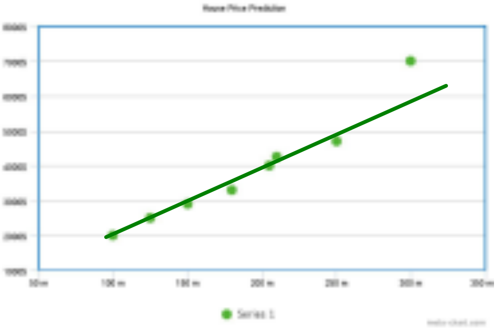


- Example scenarios:
  - Prices of house to be fitted as a function of the area of the house
  - Temperature of a place to be fitted as a function of its latitude and longitude and time of the year
  - Stock Price (or BSE/Nifty value) to be fitted as a function of Company Earnings

    Multivariate regression
  - Height of students to be fitted as a function of their weight
- One or more observations/parameters in the data are expected to represent the output in the future

Higher you go, the more expensive the house!

- Consider the variation of price (in \$) of house with variations in its height (in m) above the ground level (Mumbai)
- These are specified as coordinates of the 8 points: $(x_1, y_1), \ldots, (x_8, y_8)$
- Desired: Find a pattern or curve that characterizes the price as a function of the height



Errors and Causes

- (Observable) Data is generally collected through measurements or surveys
  - Surveys can have random human **errors** → y axis error
  - Measurements are subject to **imprecision** of the measuring or recording instrument → x-axis error
  - **Outliers** due to variability in the measurement or due to some experimental error;
- **Robustness to Errors:** Minimize the effect of error in predicted model
- **Data cleansing**: Outlier handling in a pre-processing step

→ Most often, you want the model building to do implicit data cleansing.

Curve Fitting: The Process

- *Curve fitting is the process of constructing a curve, or mathematical function, that has the* **best fit** *to a series of data points, possibly subject to constraints.* - Wikipedia

Curve Fitting: The Process

- *Curve fitting is the process of constructing a curve, or mathematical function, that has the* **best fit** *to a series of data points, possibly subject to constraints.* - Wikipedia
- Need quantitative criteria to find the **best fit**
- **Error function** E : curve $f \times$ dataset $\mathcal{D} \longrightarrow \Re$
- Error function must capture the deviation of prediction from expected value

Example

- Consider the two candidate prediction curves in blue and red respectively respectively. Which is the better fit?

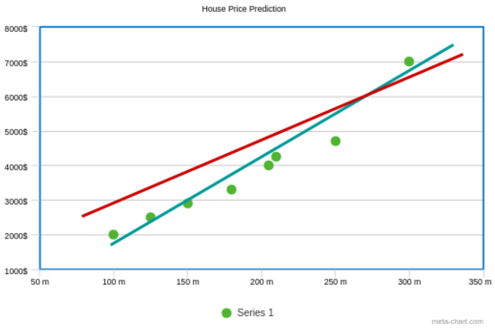
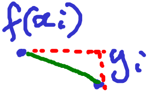


Figure: Price of house *vs.* its height - for illustration purpose only

Question



What are some options for error function $E(f, D)$ that measure the deviation of prediction from expected value?

$\sum_i (f(x_i) - y_i)^2$: Eucledean distance

$\sum_i |f(x_i) - y_i|$: Manhattan distance

$\sum_i f(x_i) - y_i$: Unsigned distance (when bias is desired)

Examples of E

- $\sum_D f(x_i) - y_i$
- $\sum_D |f(x_i) - y_i|$
- $\sum_D (f(x_i) - y_i)^2$
- $\sum_D (f(x_i) - y_i)^3$
- and many more

In 1-d case, ie $y_i, f(x_i) \in \mathbb{R}$
these behave similarly except
that $(f(x_i) - y_i)^2$ discourages
really far away pts
So $(f(x_i) - y_i)^2$ might be
less robust to
outliers than
$|f(x_i) - y_i|$

Fixing outlier sensitivity
can be through ① Fixing error fn OR
② Regularization / bayesian estimation

Question

Which choice F do you think can give us best fit curve and why?
Hint: Think of these errors as distances.

Squared Error

$$\sum_{D} (f(x_i) - y_i)^2$$

- One best fit curve corresponds to f that minimizes the above function. It..
  1. Is continuous and differentiable
  2. Can be visualized as square of Euclidean distance between predicted and observed values
- Mathematical optimization of this function: Topic of following lectures.
- This is the Method of least squares

Regression, More Formally

- Formal Definition
- Types of Regression
- Geometric Interpretation of least square solution

Linear Regression as a canonical example

- **Optimization** (Formally deriving least Square Solution)
- **Regularization** (Ridge Regression, Lasso), **Bayesian Interpretation** (Bayesian Linear Regression)
- **Non-parametric estimation** (Local linear regression),
- **Non-linearity through Kernels** (Support Vector Regression)

Linear Regression with Illustration

- Regression is about learning to predict a set of output variables (*dependent variables*) as a function of a set of input variables (*independent variables*)
- Example
  - A company wants to determine how much it should spend on T.V commercials to increase sales to a desired level y^*
  - **Basis?**

Linear Regression with Illustration

- Regression is about learning to predict a set of output variables (*dependent variables*) as a function of a set of input variables (*independent variables*)
- Example
  - A company wants to determine how much it should spend on T.V commercials to increase sales to a desired level y^*
  - **Basis?** It has previous observations of the form $<x_i, y_i>$,
    - x_i is an instance of money spent on advertisements and y_i was the corresponding observed sale figure

Linear Regression with Illustration

- Regression is about learning to predict a set of output variables (*dependent variables*) as a function of a set of input variables (*independent variables*)
- Example
  - A company wants to determine how much it should spend on T.V commercials to increase sales to a desired level y^*
  - **Basis?** It has previous observations of the form $<x_i, y_i>$,
    - x_i is an instance of money spent on advertisements and y_i was the corresponding observed sale figure
  - Suppose the observations support the following linear approximation

$$y = \beta_0 + \beta_1 * x \tag{1}$$

Then $x^* = \frac{y^* - \beta_0}{\beta_1}$ can be used to determine the money to be spent

- **Estimation** for Regression: Determine appropriate value for β_0 and β_1 from the past observations $\beta_0, \beta_1 = \arg\min \sum_{x \in D} (y - f(x))^2$

Linear Regression with Illustration

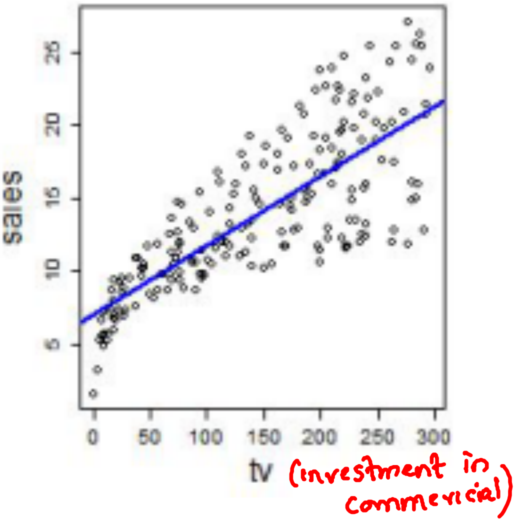


Figure: Linear regression on T.V advertising vs sales figure

What will it mean to have sales as a non-linear function of investment in advertising?

$(\text{value})^2$ $\log(\text{value})$ or time of day of advertise-ment

$$y \approx f(\phi_1(x), \phi_2(x) \,..\, \phi_n(x))$$

ϕ_i: value

x is an object representing specific investments

$\phi_1 \,..\, \phi_n$ are basis fns or features with hope that linear combination of ϕ's as f is a good approx to y

$$f(x) = w^T\phi(x)$$

[Step1 to non-linearity]

Basic Notation

- Data set: $\mathcal{D} =< \mathbf{x_1}, \mathbf{y_1} >, .., < \mathbf{x_m}, \mathbf{y_m} >$
  - Notation (used throughout the course)
  - m = number of training examples
  - $\mathbf{x}'\mathbf{s}$ = input/independent ~~variables~~ objects
  - $\mathbf{y}'\mathbf{s}$ = output/dependent/'target' variables
  - $(\mathbf{x}, \mathbf{y})$ - a single training example
  - $(\mathbf{x_j}, \mathbf{y_j})$ - specific example (j^{th} training example)
  - j is an index into the training set
- ϕ_i's are the attribute/basis functions, and let

$$\phi = \begin{bmatrix} \phi_1(\mathbf{x_1}) & \phi_2(\mathbf{x_1}) & ...... & \phi_p(\mathbf{x_1}) \\ . & & & \\ . & & & \\ \phi_1(\mathbf{x_m}) & \phi_2(\mathbf{x_m}) & ...... & \phi_p(\mathbf{x_m}) \end{bmatrix} \tag{2}$$

-

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ y_m \end{bmatrix} \tag{3}$$

Formal Definition

- **General Regression problem**: Determine a function f^* such that $f^*(x)$ is the best predictor for y, with respect to $\mathcal{D}$:

$$f^* = \underset{f \in F}{\operatorname{argmin}}\ E(f, \mathcal{D})$$

  Here, F denotes the class of functions over which the error minimization is performed

- **Parametrized Regression problem**: Need to determine parameters $\mathbf{w}$ for the function $f\left(\phi(\mathbf{x}), \mathbf{w}\right)$ which minimize our error function $E\left(f(\phi(\mathbf{x}), \mathbf{w}), \mathcal{D}\right)$

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \left\langle E\left(\underbrace{f(\phi(\mathbf{x}), \mathbf{w})}_{\text{parametrized } f}, \mathcal{D}\right) \right\rangle$$

Types of Regression

- Classified based on the function class and error function
- F is space of linear functions $f(\phi(\mathbf{x}), \mathbf{w}) = \mathbf{w}^T\phi(\mathbf{x}) + b \implies$ Linear Regression
  - Problem is then to determine $\mathbf{w}^*$ such that,

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

Types of Regression (contd.)

$\|w\|_2^2$

- **Ridge Regression:** A shrinkage parameter (regularization parameter) is added in the error function to reduce discrepancies due to variance
- **Logistic Regression:** Models conditional probability of dependent variable given independent variables and is extensively used in classification tasks

$$y \in \{0,1\} \quad \& \quad f(\phi(\mathbf{x}), \mathbf{w}) = \log \frac{\Pr(\mathbf{y}|\mathbf{x})}{1 - \Pr(\mathbf{y}|\mathbf{x})} = b + \mathbf{w}^T * \phi(\mathbf{x}) \tag{5}$$

- Lasso regression, Stepwise regression and several others

Least Square Solution

- Form of $E()$ should lead to accuracy and tractability
- 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

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

$f(x) = w^T\phi(x) + b$

- The least square solution for linear regression is obtained as

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

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

$$\Phi w^* = y$$

- 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 $\phi\mathbf{w}^* = \mathbf{y}$, where

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

and

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

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

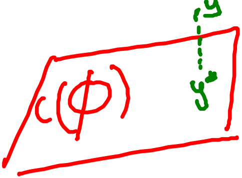
- The minimum value of the squared loss is zero
- If zero were NOT attainable at $\mathbf{w}^*$, what can be done?

When?

ϕ is not full column rank $\Rightarrow$ ① Because $n > m$ (too many irrelevant attributes)

② Redundancy in ϕ

Geometric Interpretation of Least Square Solution



$(y - y^*) \perp C(\phi)$

- 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...........

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

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

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

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

Recap: $y^* \in C(\phi) \ldots$ Let $y^* = \phi w$

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

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

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

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

Assumes

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

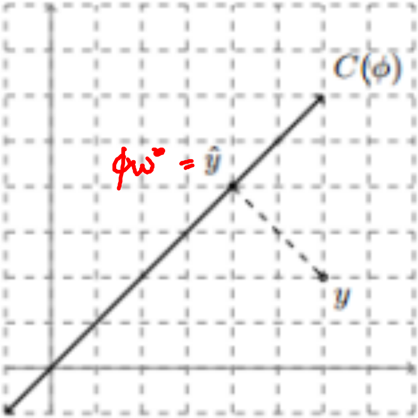
Proof?

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 $\phi^T\phi\mathbf{x} = \mathbf{0}$

ie $\phi^T\phi$ is NOT full row/column rank.

$$\Rightarrow \mathbf{x^T}\phi^\mathbf{T}\phi\mathbf{x} = \mathbf{0}$$
$$\Rightarrow (\phi\mathbf{x})^\mathbf{T}\phi\mathbf{x} = \mathbf{0} \Rightarrow \|\phi x\|_2^2 = 0$$
$$\Rightarrow \phi\mathbf{x} = \mathbf{0}$$

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 $\mathbf{x} = \mathbf{0}$ **, This implies ϕ has full column rank if $\phi^T\phi$ is invertible. Hence, theorem proved**



if $(\phi^T\phi)^{-1}$ exists then

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

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

What is Next?

- Some more questions on the Least Square Linear Regression Model
- More generally: How to minimize a function?
  - Level Curves and Surfaces
  - Gradient Vector
  - Directional Derivative
  - Hyperplane
  - Tangential Hyperplane

  → Ridge regression

  → Lasso regression
- Gradient Descent Algorithm