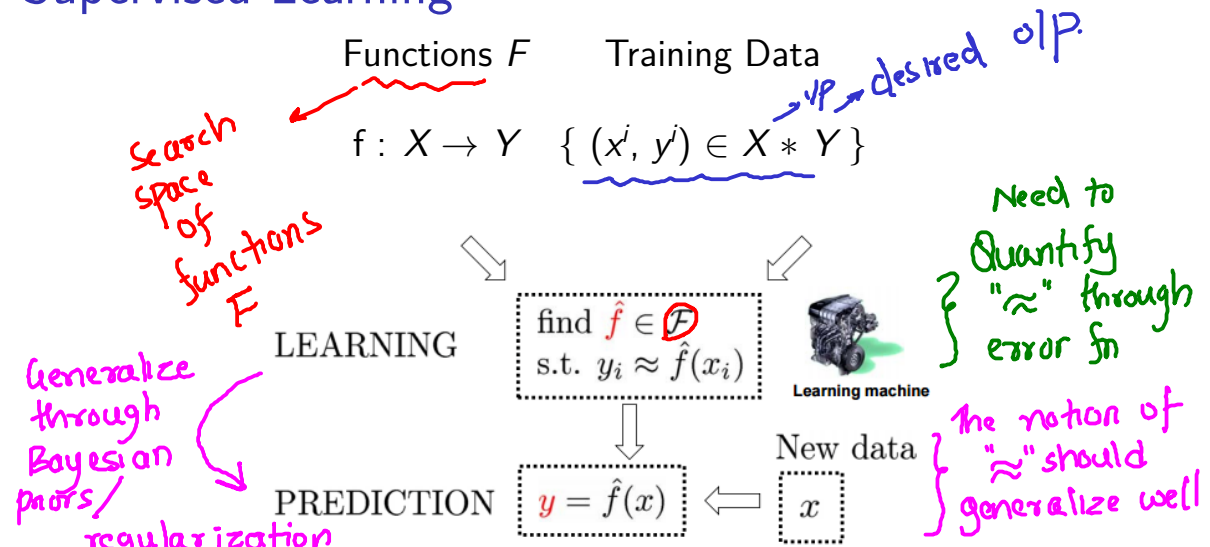
Lecture 2 - (Linear) Regression

Instructor: Prof. Ganesh Ramakrishnan

Supervised Learning



Linear Regression and Least Squares

We will start with (a) linear regression and (b) the least square method to calculate parameters for linear regression.

Our blue "best fit" line was obtained using a variant of least square.

Recap

- **Machine Learning in general**
  - Supervised Learning
  - Unsupervised Learning
  - Applications and examples
- **Canonical Learning Problems**
  - Regression Supervised
  - Classification Supervised
  - Unsupervised modeling of data

Agenda

- What is data?
  - Noise in data
- How to predict?
  - Fitting a curve
  - Error measurement
  - Minimizing Error
- Method of Least Squares

What is data?

Price of house;- Data = { locality, area, nature, facilities, mall?, .. }

- For us, data is the information about the problem, you are solving using ML, in quantized form
- This data can be from any source, some examples are
  - Prices of stock and stock indexes such as BSE or Nifty
  - Prices of house, area and size of the house
  - Temperature of a place, latitude, longitude and time of year
- The objective is to predict something using all/some of the given data
- Hence, one or more than one parameters of the data must also represent the output of our program

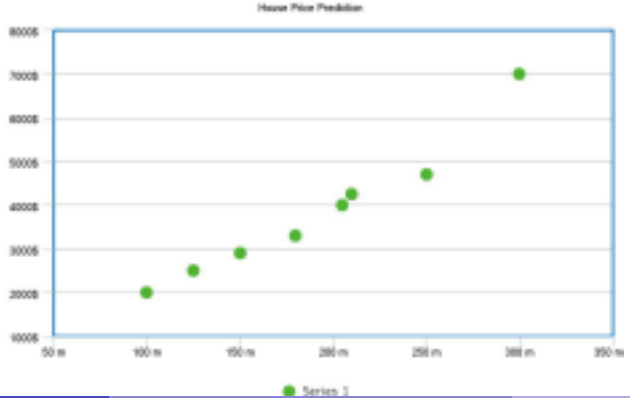
The variable to be predicted may be known for some pts & unknown for others

Noise in Data

- Data in real life problems are generally collected through surveys/measurements
- Some reasons for noise:
  - Measurements may have random human errors
  - Measurements can also be incomplete
- One aspect of data cleansing is outlier detection

Example dataset for this lecture

- Consider variation of cost of the house with floor/height
- Find a pattern or curve which this dataset follows, hence predict the price for any floor



How to predict?

- 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
- Thus we need a critera to compare two curves on a dataset
- We describe an error function $E(f, D)$ which takes a curve f and dataset D as input and returns a real number
- Error function must be such that it can capture how bad the prediction is

Error(prediction fn, Data set)
captures "$\approx$"

Example

* Robustness ⇒ Provide robustness while predicting on unseen data

- Consider the example below where we have two curves on our dataset defined by blue(f_b) and red(f_r) line respectively. We want to find which is the better fit.

Desirables:
1) $E \to 0$ as $y_i \pm f(x_i)$
2) $E \geq 0$ (ideal state)
3) Robust* to transformations in x_i's such as unit (x_i in floor #, in feet / yards)

f_b

f_r

4) Differentiable wrt function/ model params (so that derivative can be set to 0)

Question

Robustness is not only a fn of $E(f, D)$ but also of f & algorithm used to minimize $E(f, D)$

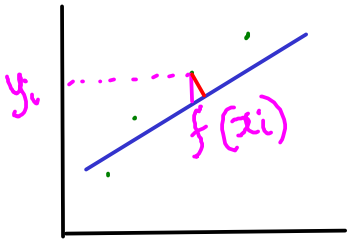
$\frac{1}{m}\sum_i |y_i - f(x_i)|$ 3/4

$\frac{1}{m}\sum_i (y_i - f(x_i))^2$ 4/4

$\sqrt{\frac{1}{m}\sum_i (y_i - f(x_i))^2}$ 4/4

What are some options for $E(f, D)$?

Hint: Measurement of difference from original value.



$\frac{1}{m}\sum_i \frac{\perp \text{ distance of } (x_i, y_i)}{\text{from } f(\cdot)}$ 3/4

$\frac{1}{m}\sum \left[(y_i - f(x_i))/|f(x_i)|\right]^2$

You might want to put such scalings into the model instead of the error function

Examples of E

$f(x_i) \to -\infty$ will decrease $E \to -\infty$

- $\sum_D f(x_i) - y_i$ 1/4
- $\sum_D |f(x_i) - y_i|$ 3/4
- $\sum_D (f(x_i) - y_i)^2$ 4/4 $\rightarrow$ $\sum_D (f(x_i) - y_i)^{2k}$ $k = 1 \ldots$
- $\sum_D (f(x_i) - y_i)^3$ 1/4
- and many more

Boosting style algos give more importance to pts with larger errors.

Question

What E do you think can give us best fit curve and why?
Hint: Intuition of distances.

$$E(f, D) = \sum_{i=1}^{m} \left(f(x_i) - y_i\right)^2$$

Method of Least Squares

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

- To find the best fit curve we try to minimize the above function
- It is continuous and differentiable
- It can be visualized as square of Euclidean distance between predicted points and actual points
- Mathematical treatment of this function will be covered subsequently

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)

Varying form of E

→ Varying form of f

Varying form of f & E

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?**

$\phi(x_i) = [\phi_1(x_i) \quad \phi_2(x_i) \quad \ldots \quad \phi_k(x_i)]$

(x_i ≠ (commercial #i))

$\phi_1(x_i)$ Duration

$\phi_2(x_i)$ Cost of prod

$\phi_k(x_i)$ Target age

y_i = sale

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

$\phi_i(x_i)$ = amount spent on x_i

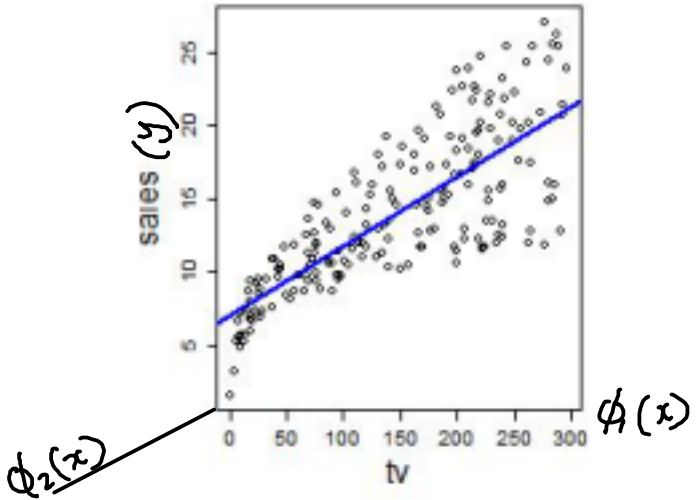
Suppose the observations support the following linear approximation

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

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

Linear Regression with Illustration



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

$$y_i = w_1\phi_1(x_i) + w_2\phi_2(x_i) + w_3\phi_1(x_i)\phi_2(x_i) + w_4\phi_1^2(x_i) \cdots$$

$\phi_3(x_i)$

$\phi_4(x_i)$

y_i is nonlinear in $(\phi_1 \cdots \phi_k)$

y_i is linear in $(w_1, w_2 \cdots w_k \cdots)$

Linear Regression

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
    - $\mathbf{y}'\mathbf{s}$ = 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

$$\Phi = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & ..... & \phi_p(\mathbf{x}_1) \\ & . & & \\ & . & & \\ \phi_1(\mathbf{x}_m) & \phi_2(\mathbf{x}_m) & ..... & \phi_p(\mathbf{x}_m) \end{bmatrix} \quad (2)$$

- $\mathbf{y} = \begin{bmatrix} y_1 \end{bmatrix}$

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 $E\left(f(\phi(\mathbf{x}), \mathbf{w}), \mathcal{D}\right)$

LR : $f(\phi(x), w) = w^T \phi(x) = [\ \ ]$ $[\ \ ]$

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

Types of Regression

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

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

Can avoid b by: $\phi(x) = \begin{bmatrix} \phi_1 \\ \phi_2 \\ 1 \end{bmatrix}$ $w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ b \end{bmatrix}$

Types of Regression (contd.)

- **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

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

- 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

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

- The least square solution for linear regression is obtained as

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

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