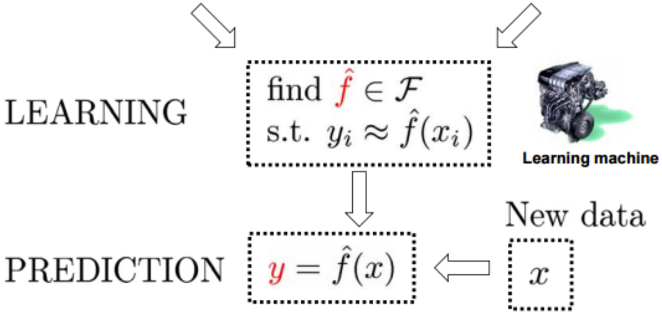
Lecture 2 - Regression

Instructor: Prof. Ganesh Ramakrishnan

Supervised Learning

Functions F Training Data

$f : X \to Y \quad \{ (x^i, y^i) \in X * Y \}$



Next ....

We will start with linear regression and least square method to calculate parameters for linear regression problems.

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?

- 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 → Pred stock value for a company in the future
  - Prices of house, area and size of the house
  - Temperature of a place, latitude, longitude and time of year
- The objective of ML is to predict or classify something using the given data
- Hence, one or more than one parameters of the data must also represent the output of our program

Noise in Data

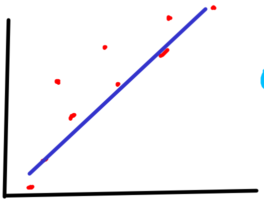
→ A sophisticated way of talking about some unaccounted measure-ment

① Light intensity without reporting exact coordinate

② weight without demographic info

- Data in real life problems are generally collected through surveys
- Surveys may have random human errors (especially qualitative)
- Most methods we will be using deal with expectations as they minimize the effect of error in our predictions

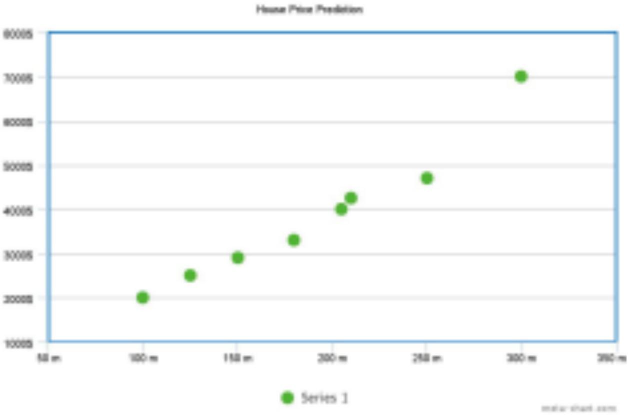
  at model level → $E(Y)$ should be as good as possible!
- Data Cleansing by finding outliers



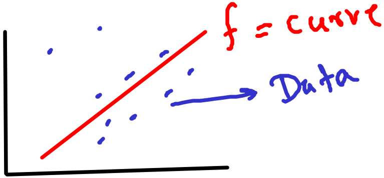
Treat as outlier

Example dataset for this lecture

- For this lecture we will consider variation of cost of the house with the area of the house
- In this example we want to find a pattern or curve which this dataset follows, hence predict the price for any value of area



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

Example

- 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.

Figure: House purchase data curve fit

Question

What are some options for $E(f, D)$?
Hint: Measurement of difference from original value.

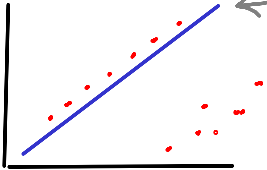
Examples of E

Non-differentiable

Assymmetric

Heavily penalize large deviations

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



With assymmetric E one side is left out

Question

Eucledian space is a vector space

V is v.s on $\mathbb{R}$ if

$\forall v_1, v_2 \in V, \ \alpha_1 v_1 + \alpha_2 v_2 \in V$

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

<u>Note:</u> $\sum_{D} (f(x_i) - y_i)^2$ = squared distance in Eucledian space $\left(\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \right)$

$= \|f - y\|_2^2$

Squared Error

$$\sum_{D} (f(x_{\mathrm{i}}) - y_{\mathrm{i}})^2$$

- 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
- How we can perform mathematical treatment over this function will be covered in further lectures.
- This mathematical treatment is known as 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

Use of Linear regression

- 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

money to spend

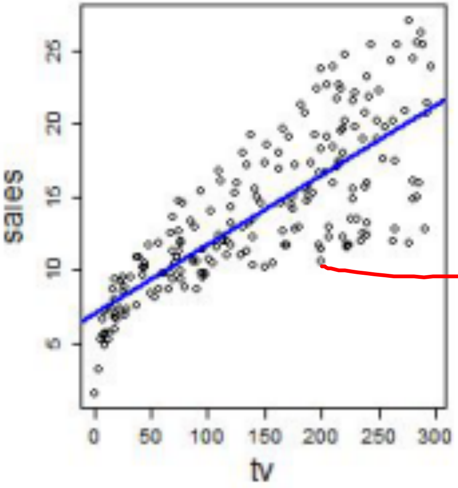
Observed Sale

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

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



Bayesian Linear Regression

Prediction based on far away pts might have lower confidance

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?

What can be non-linear about linear regression?

Ans: f is nonlinear in x

linear in ω

Thru $\phi(x)$

$\phi_1(x)$ = time of day $\phi_2(x)$ = the cost - - - -

Basic Notation

some object

Eg: Advertisement possibility } x

Time of day

Avenue

Company

ϕ_1 ϕ_2 ϕ_3

- 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

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

-

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ y_m \end{bmatrix} \quad (3)$$

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

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

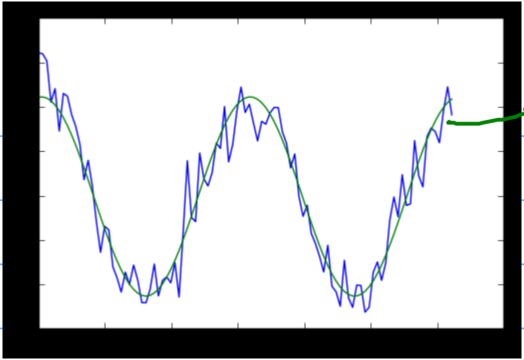
$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \Big\langle E\big(f(\phi(\mathbf{x}), \mathbf{w}), \mathcal{D}\big) \Big\rangle$$

$f^* \approx f(\ldots w^*)$

Types of Regression

- Classified based on the function class and error function
- E 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,

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



$f = w_1^* + w_2^* x + w_3^* x^2$

$+ \cdots$

$w^* = [\ 1.96075676 \quad 0.40288363 \quad 4.46959679]$

$\phi = [\quad 1 \qquad x \qquad x^2$

Types of Regression (contd.)

Modify objective of linear regression

- **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

transforms "y"

$y \in \{0,1\}$

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

- 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

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

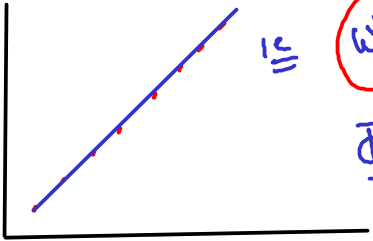
- The least square solution for linear regression is obtained as

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

- The minimum value of the squared loss is zero
- If zero were attained at $\mathbf{w}^*$, we would have ......$f(x_i) = y_i$...... $\forall i$

is $w^T \phi(x_i) + b = y_i \ \forall i$

is compactly $\Phi w = Y$



$$\Phi = \begin{bmatrix} \phi(x_1) & 1 \\ \vdots & \\ \phi(x_m) & 1 \end{bmatrix} \quad w = \begin{bmatrix} w_1 \\ \vdots \\ w_p \\ b \end{bmatrix}$$

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

$y_i \neq f(x_i)$ for each i

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............

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

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

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

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

Nearly dependent: characterized through condition # of $\Phi^T\Phi$

$\phi_k(\cdot) \approx \sum_j \phi_j(\cdot)\,\alpha_j$

(cannot be detected through rank)

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

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

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

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

- Here $\Phi^T\Phi$ is invertible if and only if Φ has full column rank

Features must be independent!

Problem: In several engineering setups, features are designed to be correlated

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 $\Phi\mathbf{x} = 0 \Rightarrow \mathbf{x} = 0$

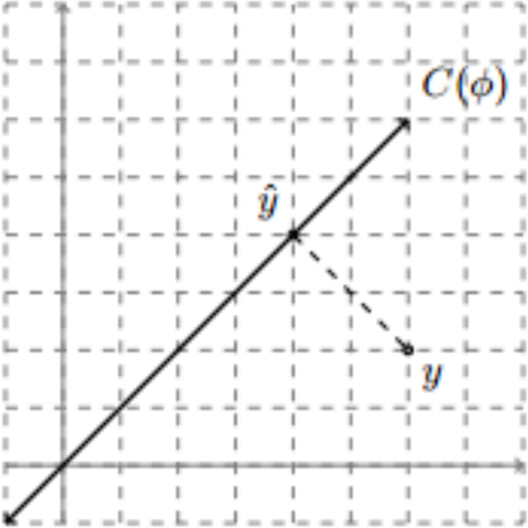
Assume on the contrary that $\Phi^T\Phi$ is non invertible. Then $\exists\mathbf{x} \neq 0$ such that $\Phi^T\Phi\mathbf{x} = 0$

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

$\|\phi x\|_2^2 = 0$

This is a contradiction. Hence $\Phi^T\Phi$ is invertible if Φ is full column rank

If $\Phi^T\Phi$ is invertible then $\Phi\mathbf{x} = 0$ implies $(\Phi^T\Phi\mathbf{x}) = 0$, which in turn implies $\mathbf{x} = 0$, This implies Φ has full column rank if $\Phi^T\Phi$ is invertible. The converse can also be proved similarly.


$C(\phi)$
$\hat{y}$
y

How about an Analytic Derivation?

- 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

  Black box techniques for minimizing.

- Gradient Descent Algorithm