Lecture 2, Part 2 - Regression

Instructor: Prof. Ganesh Ramakrishnan

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

Linear Regression with Illustration

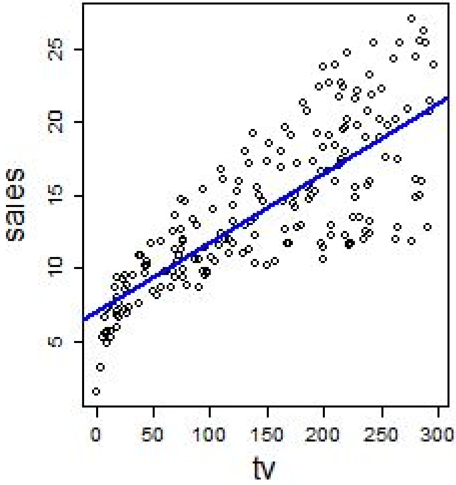

Figure: Linear regression on T.V advertising vs sales figure

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

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

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

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

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 $f\big(\phi(\mathbf{x}), \mathbf{w}\big)$ which minimize our error function $E\big(f(\phi(\mathbf{x}), \mathbf{w}), \mathcal{D}\big)$

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

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,

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

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

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

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

- 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?

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} &\Rightarrow \mathbf{x}^T\Phi^T\Phi\mathbf{x} = 0 \\ &\Rightarrow (\Phi\mathbf{x})^T\Phi\mathbf{x} = 0 \\ &\Rightarrow \Phi\mathbf{x} = 0 \end{aligned}$$

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 x = 0 , This implies Φ has full column rank if $\Phi^T\Phi$ is invertible. The converse can also be proved similarly.

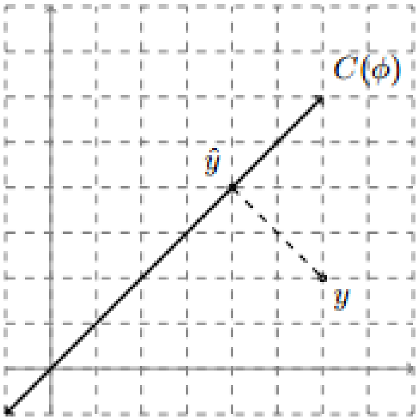

Figure: Least square solution $\mathbf{y}^*$ is the orthogonal projection of y onto column space of Φ

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
- Gradient Descent Algorithm