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
- Input variables are called *independent variables*
- Output variables are called *dependent variables*

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

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

  and then find the money they need to spend using this function
- Regression problem is to find the appropriate function and its coefficients

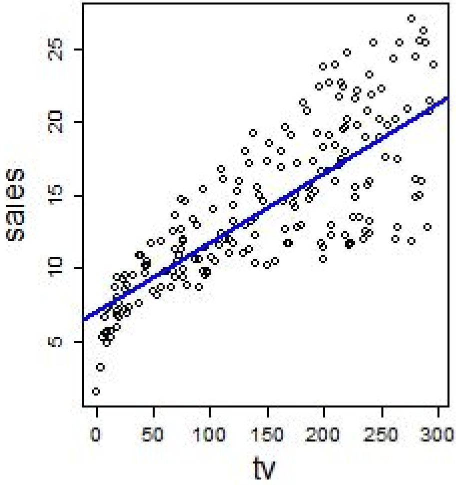

Figure: Linear regression on T.V advertising vs sales figure

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

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 $y_i \; \forall i$
- Regression problem: Determine a function f^* such that $f^*(x)$ is the best predictor for y, with respect to D,

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

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
  - 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
- Logistic Regression : Used to model conditional probability of dependent variable given independent variable and is extensively used in classification tasks

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

- Lasso regression, Stepwise regression and many more

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

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

- 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 $\forall u, \phi^T(x_u)\mathbf{w}^* = \mathbf{y_u}$, 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?

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

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

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

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

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

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

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

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

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

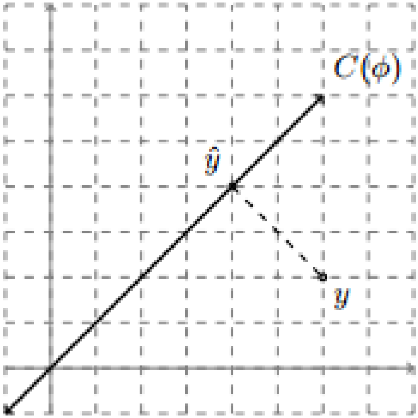

Figure: Least square solution $\mathbf{y}^*$ is the orthogonal projection of y onto column space of ϕ