Introduction to Machine Learning - CS725

Instructor: Prof. Ganesh Ramakrishnan

Lecture 4 - Least Squares Linear Regression

Regression Model

- Training set (this is your data set), $\mathcal{D} =< \mathbf{x_1}, \mathbf{y_1} >, < \mathbf{x_2}, \mathbf{y_2} >, .., < \mathbf{x_m}, \mathbf{y_m} >$
  - Notation (used throughout the course)
  - m = number of training examples
  - $\mathbf{x's}$ = input variables / features
  - $\mathbf{y's}$ = output variable "target" variables
  - $(\mathbf{x}, \mathbf{y})$ - single training example
  - $(\mathbf{x_i}, \mathbf{y_i})$ - specific example (ith training example)
  - i is an index to training set
- Need to determine parameters $\mathbf{w}$ for the function $f(\mathbf{x}, \mathbf{w})$ which minimizes our error function $\varepsilon(f(\mathbf{x}, \mathbf{w}), \mathcal{D})$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\{ \varepsilon(\mathbf{f}(\mathbf{x}, \mathbf{w}), \mathcal{D}) \right\}$$

Linear Regression Model

- Need to determine $\mathbf{w}$ for the linear function $\underline{f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{p} w_i \phi_i(\mathbf{x_j}) = \phi \mathbf{w}}$ which minimizes our error function $\underline{\varepsilon(f(\mathbf{x}, \mathbf{w}), \mathcal{D})}$ → unspecified

specified

- ϕ_i's are the basis functions, and let

$$\phi = \begin{bmatrix} \phi_1(\mathbf{x_1}) & \phi_2(\mathbf{x_1}) & \ldots\ldots & \phi_p(\mathbf{x_1}) \\ . & & & \\ . & & & \\ \phi_1(\mathbf{x_m}) & \phi_2(\mathbf{x_m}) & \ldots\ldots & \phi_p(\mathbf{x_m}) \end{bmatrix} \quad (1)$$

- $$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_m \end{bmatrix} \quad (2)$$

Least Square Linear Regression Model

- $$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ w_p \end{bmatrix} \quad (3)$$

- $$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left\{ \sum_{\mathbf{j=1}}^{\mathbf{m}} \left(\sum_{\mathbf{i=1}}^{\mathbf{p}} \mathbf{w_i} \phi_{\mathbf{i}}(\mathbf{x_j}) - \mathbf{y_j} \right)^{\mathbf{2}} \right\} \quad (4)$$

- $$\varepsilon = \min_{\mathbf{w}} \left(\mathbf{w^T} \phi^{\mathbf{T}} \phi \mathbf{w} - \mathbf{2y^T} \phi \mathbf{w} + \mathbf{y^T y} \right) \quad (5)$$

Form of error fn also specified

Value of error

Recap

- **Regression**
  - Formal Definition
  - Examples and Types of Regression
- **Least Square Solution**
  - Role of error/loss function
  - Least square solution for linear regression
- **Geometric Interpretation of Least Square Solution**
- **Theorem** : $\phi^T\phi$ is invertible if and only if ϕ is full column rank

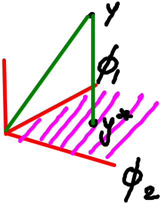
Geometric Interpretation of Least Square Solution

- 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

in column space of ϕ

To be predicted

- Therefore, the line joining $\mathbf{y}^*$ to $\mathbf{y}$ should be orthogonal to the column space



$$\phi \mathbf{w} = \mathbf{y}^* \tag{6}$$

$$(\mathbf{y} - \mathbf{y}^*)^{\mathbf{T}} \phi = \mathbf{0} \tag{7}$$

$$(\mathbf{y}^*)^{\mathbf{T}} \phi = (\mathbf{y})^{\mathbf{T}} \phi \tag{8}$$

$$(\phi\mathbf{w})^{\mathbf{T}}\phi = \mathbf{y}^{\mathbf{T}}\phi \tag{9}$$

$$\mathbf{w}^{\mathbf{T}}\phi^{\mathbf{T}}\phi = \mathbf{y}^{\mathbf{T}}\phi \tag{10}$$

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

$$\mathbf{w} = (\phi^{\mathbf{T}}\phi)^{-1}\phi^{T}y \tag{12}$$

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

Theorem : $\phi^T\phi$ is invertible if and only if ϕ is full column rank

Proof :

only if

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

Proof by contradiction

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

$a^Ta = 0$ iff $a = 0$

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

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

This is a contradiction. Hence $\phi^T\phi$ is invertible if ϕ is full column rank

if

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**

Agenda

- 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

Some questions

If A is p.d then all $\lambda(A) > 0 \Rightarrow Ax = 0$ has only $x = 0$ as solution, because if $\exists x \neq 0$ s.t $Ax = 0$ then $\lambda = 0$ will be an eigenvalue of A (ie $Ax = \lambda x$)
$\Rightarrow A$ must be invertible ... So suffices to check that $\phi^T\phi$ has no zero eigenvalues

- What is the relationship between positive definiteness and invertibility? $\rightarrow x^T(\phi^T\phi)x = \|\phi x\|_2^2 \geq 0 \Rightarrow$ All $\lambda(\phi^T\phi) \geq 0$
- When is ϕ not full column rank? What are associated problems and fixes? $\rightarrow$ Eg: If $m < p$, $\phi_i(x) = \phi_j(x)\ \forall x$ OR $\sum_{t=1}^{k} \phi_{i_t}(x) = \sum_{t=1}^{k} \phi_{j_t}(x)$
- How to find a solution if ϕ is not full column rank?

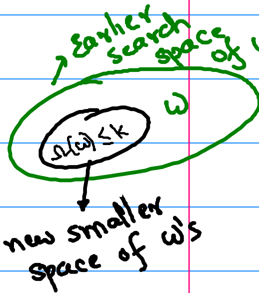
Feature selection
① Select only a subset of ϕ_i's & drop the rest st subset is linearly independent
Problem: 2^n subsets to explore !!
Algos for greedily selecting ϕ_i to include or exclude
Eg: Infogain (ϕ_i | attributes selected so far) followed by Least squares soln

② Modify the objective being minimized

Modifying objective?

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

$$\text{s.t.} \quad \Omega(w) \le k$$



$$= \underset{}{\operatorname{argmin}} \; \|\phi w - y\|^2$$

$$\text{s.t.} \quad \Omega(w) \le k$$

Eg: $\Omega(w) = \#$ of non-zero w_i's

ie Least squares regression s.t not more than k w's are $\neq 0$ ie not more than k ϕ_i's that are "effective"

3 solns $\rightarrow$ (1) feature selection on ϕ's to give ϕ_s, followed by $w^* = (\phi_s^T \phi_s)^{-1} \phi_s^T y$

(2) $w^* = \arg\min_{w} \|\phi w - y\|^2$

s.t. $\Omega(w) \leq \xi$

(3) $\phi^T \phi w^* = \phi^T y$

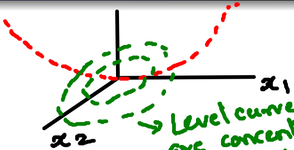
H/w problem: Based on diff inequalities between p & m, find cases where this equation has (a) No solution (b) One solution and (c) multiple solutions

Solving Least Square Linear Regression Model

- Intuitively: Minimize by setting derivative (gradient) to 0 and find closed form solution.
- For most optimization problems, finding closed form solution is difficult
- Even for linear regression (for which closed form solution exists), are there alternative methods?
- Eg: Consider, $\mathbf{y} = \phi\mathbf{w}$,where ϕ is a matrix with full column rank, the least squares solution, $\mathbf{w}^* = \phi^T\phi)^{-1}\phi^T\mathbf{y}$. Now, imagine that ϕ is a very large matrix. with say, 100,000 columns and 1,000,000 rows. Computation of closed form solution might be challenging.
- How about an iterative method?

Level curves and surfaces

Eg: $f(x_1, x_2) = x_1^2 + x_2^2$



Level curves are concentric circles on x_1, x_2 plane

- A level curve of a function $\mathbf{f}(\mathbf{x})$ is defined as a curve along which the value of the function remains unchanged while we change the value of it's argument x.
- Formally we can define a level curve as :

$$L_c(\mathbf{f}) = \left\{ \mathbf{x} | \mathbf{f}(\mathbf{x}) = \mathbf{c} \right\} \tag{13}$$

where c is a constant.

Level curves and surfaces

- The image below is an example of different level curves for a single function

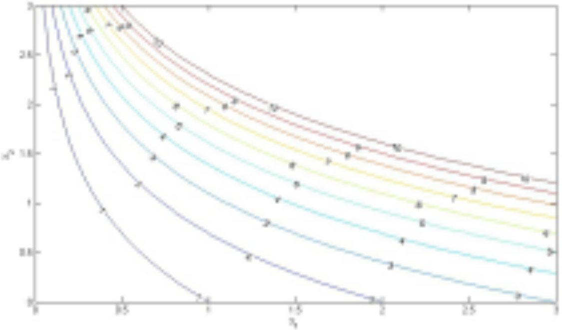


Figure 1: 10 level curves for the function $\mathbf{f(x_1, x_2) = x_1 e^{x_2}}$ (Figure 4.12 from `https://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf`)

Directional Derivatives

- Directional derivative: Rate at which the function changes at a given point in a given direction
- The *directional derivative* of a function f in the direction of a unit vector $\mathbf{v}$ at a point $\mathbf{x}$ can be defined as :

$$D_{\mathbf{v}}(f) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{v}) - \mathbf{f}(\mathbf{x})}{h} \tag{14}$$

$$||\mathbf{v}|| = \mathbf{1} \tag{15}$$

Distance traversed along v

v is treated as a unit vector

Gradient Vector

∇f is direction of maximum directional derivative

- Magnitude (euclidean norm) of gradient vector at any point indicates maximum value of directional derivative at that point
- Direction of gradient vector indicates direction of this maximal directional derivative at that point.
- The *gradient vector* of a function f at a point $\mathbf{x}$ is defined as:

$$\nabla f_{\mathbf{x}^*} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ . \\ . \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \epsilon \mathbb{R}^n \tag{16}$$

$D_v(f(x)) = \nabla^T f(x)\, v$

$\Downarrow$

By Cauchy Schwarz ineq, $D_v(f(x))$ is max when v is in the direction of $\nabla f(x)$

Gradient Vector

- Magnitude (euclidean norm) of gradient vector at any point indicates maximum value of directional derivative at that point
- The *gradient vector* of a function f at a point $\mathbf{x}$ is defined as:

$$\nabla f_{\mathbf{x}^*} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ . \\ . \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \epsilon \mathbb{R}^n \tag{17}$$

- Thus, at the point of minimum of a differentiable minimization objective (such as least squares for regression), ...

Necessary: $\nabla E(w^*) = 0$. Need to verify that by solving this eqn for least squares regression, we get $w^* = (\phi^T\phi)^{-1}\phi^T y$

Gradient Vector

- The figure below gives an example of gradient vector

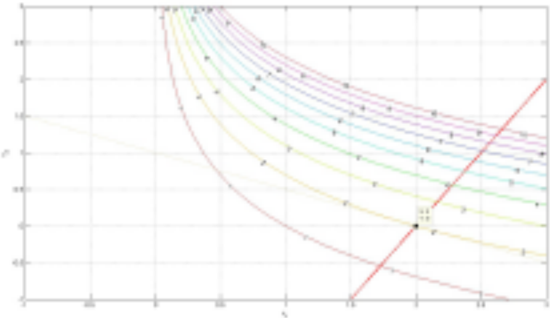


Figure 2: The level curves from Figure 1 along with the gradient vector at (2, 0). Note that the gradient vector is perpenducular to the level curve $x_1 e^{x_2} = 2$ at (2, 0)