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 ($\mathbf{i}^{\text{th}}$ 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 $f(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{p} w_i \phi_i(\mathbf{x_j}) = \phi \mathbf{w}$ which minimizes our error function $\varepsilon(f(\mathbf{x}, \mathbf{w}), \mathcal{D})$
- ϕ_i's are the basis functions, and let

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

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

Least Square Linear Regression Model

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

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

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

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

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

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

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

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

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

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

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

- 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 :
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 $\mathbf{x = 0}$ **, 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

- What is the relationship between positive definiteness and invertibility?
- When is ϕ not full column rank? What are associated problems and fixes?
- How to find a solution if ϕ is not full column rank?

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

- 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

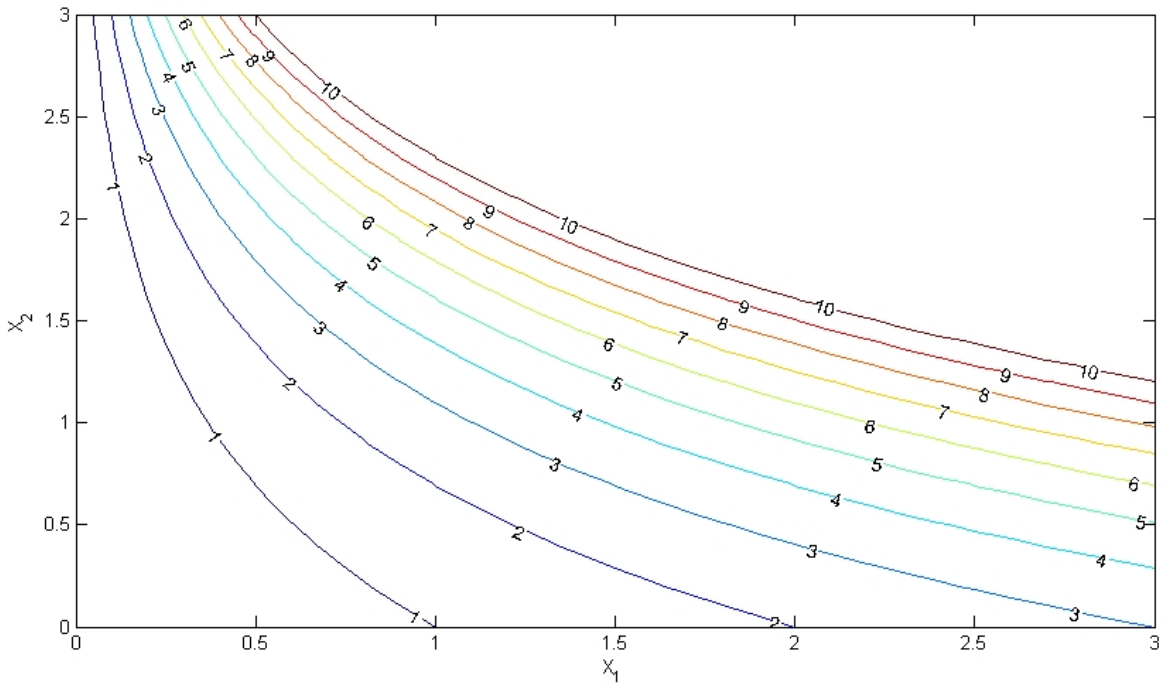

Figure 1: 10 level curves for the function $\mathbf{f}(\mathbf{x_1}, \mathbf{x_2}) = \mathbf{x_1 e_2^x}$ (Figure 4.12 from `https://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf`)

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 :

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

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

Gradient Vector

- 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:

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

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), ...

Gradient Vector

- The figure below gives an example of gradient vector

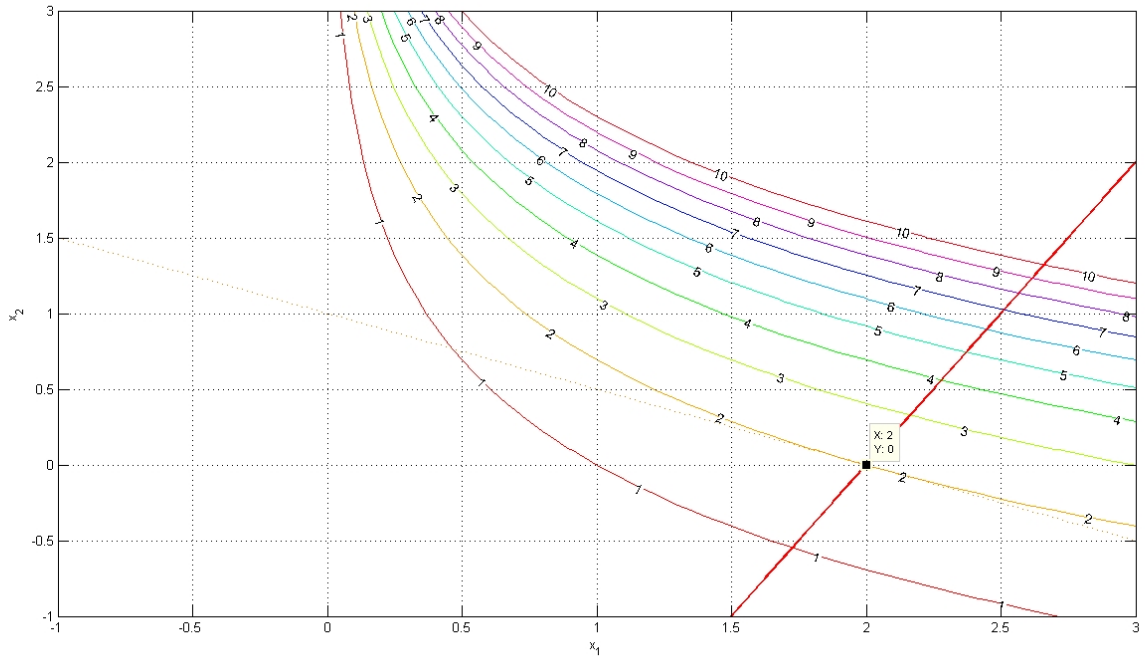

Figure 2: The level curves from Figure 1 along with the gradient vector at (2, 0). Note that the gradient vector is perpenducular to the level curve $x_1 e^{x_2} = 2$ at (2, 0)

Hyperplanes

- A hyperplane in an n-dimensional Euclidean space is a flat, n-1 dimensional subset of that space that divides the space into two disconnected parts.
- Technically, a hyperplane is a set of points whose direction *w.r.t.* a point $\mathbf{p}$ is orthogonal to a vector $\mathbf{v}$.
- Formally:

$$H_{\mathbf{v},\mathbf{p}} = \left\{ \mathbf{q} \mid (\mathbf{p} - \mathbf{q})^{\mathsf{T}} \mathbf{v} = \mathbf{0} \right\} \tag{18}$$

Tangential Hyperplanes

There are two definitions of *tangential hyperplane* ($TH_{\mathbf{x}^*}$) to *level surface* ($L_{f(\mathbf{x}^*)}(f)$) of f at $\mathbf{x}^*$:

- Plane consisting of all tangent lines at $\mathbf{x}^*$ to any parametric curve $c(t)$ on level surface.
- Plane orthogonal to the gradient vector at $\mathbf{x}^*$.

$$TH_{\mathbf{x}^*} = \left\{ \mathbf{p} \mid (\mathbf{p} - \mathbf{x}^*)^{\mathbf{T}} \nabla \mathbf{f}(\mathbf{x}^*) = \mathbf{0} \right\} \tag{19}$$

Gradient Descent Algorithm

Gradient descent is based on the observation that if the multi-variable function F($\mathbf{x}$) is defined and differentiable in a neighborhood of a point $\mathbf{a}$, then F($\mathbf{x}$) decreases fastest if one goes from $\mathbf{a}$ in the direction of the negative gradient of F at $\mathbf{a}$,i.e. -∇ F($\mathbf{a}$).
Therefore,

$$\Delta \mathbf{w^{(k)}} = - \nabla \varepsilon(\mathbf{w^{(k)}}) \qquad \textbf{from equation (5)}$$

Hence,

$$\mathbf{w^{(k+1)}} = \mathbf{w^{(k)}} + \mathbf{2t^{(k)}}(\phi^{\mathbf{T}}\mathbf{y} - \phi^{\mathbf{T}}\phi\mathbf{w^{(k)}}) \qquad (20)$$

Gradient Descent Algorithm

Find starting point $\mathbf{w^{(0)}} \epsilon \mathcal{D}$

- $\Delta \mathbf{w^k} = -\nabla \varepsilon(\mathbf{w^{(k)}})$
- Choose a step size $t^{(k)} > 0$ using exact or backtracking ray search.
- Obtain $\mathbf{w^{(k+1)}} = \mathbf{w^{(k)}} + \mathbf{t^{(k)}} \mathbf{\Delta w^{(k)}}$.
- Set $k = k + 1$. **until** stopping criterion (such as $\| \nabla \varepsilon(\mathbf{x^{(k+1)}}) \| \leq \epsilon$) is satisfied

Gradient Descent Algorithm

Exact line search algorithm to find $t^{(k)}$

- The line search approach first finds a descent direction along which the objective function f will be reduced and then computes a step size that determines how far **x** should move along that direction.
- In general,

$$t^{(k)} = \arg\min_{t} f\left(\mathbf{w^{(k+1)}}\right) \tag{21}$$

- Thus,

$$t^{(k)} = \arg\min_{t}\left(\mathbf{w^{(k)}} + \mathbf{2t}\left(\phi^{\mathbf{T}}\mathbf{y} - \phi^{\mathbf{T}}\phi\mathbf{w^{(k)}}\right)\right) \tag{22}$$

Example of Gradient Descent Algorithm

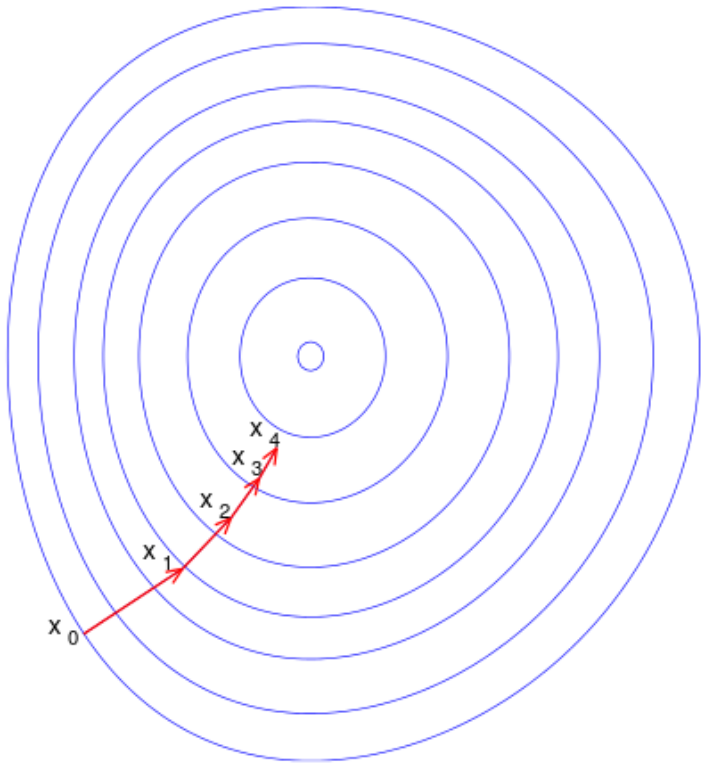

Figure 3: A red arrow originating at a point shows the direction of the negative gradient at that point. Note that the (negative) gradient at a point is orthogonal to the level curve going through that point. We see that gradient descent leads us to the bottom of the bowl, that is, to the point where the value of the function F is minimal. Sources: Wikipidea