

Introduction to Machine Learning - CS725

Instructor: Prof. Ganesh Ramakrishnan

Lecture 5a - Least Squares Linear Regression

Recall

We recall that the problem was to find \mathbf{w} such that

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \|\phi \mathbf{w} - \mathbf{y}\|^2 \quad (1)$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}} (\mathbf{w}^T \phi^T \phi \mathbf{w} - 2\mathbf{w}^T \phi \mathbf{y} - \mathbf{y}^T \mathbf{y}) \quad (2)$$

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:

$$\nabla f_{\mathbf{x}^*} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \cdot \\ \cdot \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n \quad (3)$$

- 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

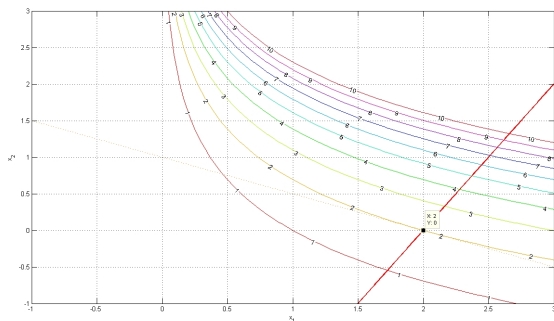


Figure: The level curves along with the gradient vector at $(2, 0)$. Note that the gradient vector is perpendicular to the level curve $x_1 e^{x_2} = 2$ at $(2, 0)$

Recall

We recall that the problem was to find \mathbf{w} such that

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \|\phi \mathbf{w} - \mathbf{y}\|^2 \quad (4)$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}} (\mathbf{w}^T \phi^T \phi \mathbf{w} - 2\mathbf{w}^T \phi \mathbf{y} - \mathbf{y}^T \mathbf{y}) \quad (5)$$

Necessary condition 1

- If $\nabla f(\mathbf{x}^*)$ is defined & \mathbf{x}^* is local minimum/maximum, then $\nabla f(\mathbf{x}^*) = 0$ (A necessary condition) (Cite : Theorem 60)¹
- Given that

$$f(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}}(\mathbf{w}^T \phi^T \phi \mathbf{w} - 2\mathbf{w}^T \phi \mathbf{y} - \mathbf{y}^T \mathbf{y})$$

$\implies \dots\dots\dots$

- we would have

$\dots\dots\dots$

$\implies \dots\dots\dots$

$\implies \dots\dots\dots$

¹convexopt.

Necessary condition 1

- If $\nabla f(\mathbf{x}^*)$ is defined & \mathbf{x}^* is local minimum/maximum, then $\nabla f(\mathbf{x}^*) = 0$ (A necessary condition) (Cite : Theorem 60)²
- Given that

$$f(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} (\mathbf{w}^T \phi^T \phi \mathbf{w} - 2\mathbf{w}^T \phi \mathbf{y} - \mathbf{y}^T \mathbf{y}) \quad (6)$$

$$\implies \nabla f(\mathbf{w}) = 2\phi^T \phi \mathbf{w} - 2\phi^T \mathbf{y} \quad (7)$$

- we would have

$$\nabla f(\mathbf{w}^*) = 0 \quad (8)$$

$$\implies 2\phi^T \phi \mathbf{w}^* - 2\phi^T \mathbf{y} = 0 \quad (9)$$

$$\implies \mathbf{w}^* = (\phi^T \phi)^{-1} \phi^T \mathbf{y} \quad (10)$$

Necessary Condition 2

- Is $\nabla^2 f(\mathbf{w}^*)$ positive definite ?
i.e. $\forall \mathbf{x} \neq 0$, is $\mathbf{x}^T \nabla^2 f(\mathbf{w}^*) \mathbf{x} > 0$? (A sufficient condition for local minimum)
(Note : Any positive definite matrix is also positive semi-definite)
(Cite : Section 3.12 & 3.12.1)³

.....
 \implies
.....
.....

- And if ϕ **has full column rank** ,

.....

\therefore If $\mathbf{x} \neq 0$, $\mathbf{x}^T \nabla^2 f(\mathbf{w}^*) \mathbf{x} > 0$

³cs709/notes/LinearAlgebra.pdf

Necessary Condition 2

- Is $\nabla^2 f(\mathbf{w}^*)$ positive definite ?
i.e. $\forall \mathbf{x} \neq 0$, is $\mathbf{x}^T \nabla^2 f(\mathbf{w}^*) \mathbf{x} > 0$? (A sufficient condition for local minimum)
(Note : Any positive definite matrix is also positive semi-definite)
(Cite : Section 3.12 & 3.12.1)⁴

$$\nabla^2 f(\mathbf{w}^*) = 2\phi^T \phi \quad (11)$$

$$\implies \mathbf{x}^T \nabla^2 f(\mathbf{w}^*) \mathbf{x} = 2\mathbf{x}^T \phi^T \phi \mathbf{x} \quad (12)$$

$$= 2(\phi \mathbf{x})^T \phi \mathbf{x} \quad (13)$$

$$= 2\|\phi \mathbf{x}\|^2 \geq 0 \quad (14)$$

- And if ϕ **has full column rank** ,

$$\phi \mathbf{x} = 0 \quad \text{iff} \quad \mathbf{x} = 0 \quad (15)$$

\therefore If $\mathbf{x} \neq 0$, $\mathbf{x}^T \nabla^2 f(\mathbf{w}^*) \mathbf{x} > 0$

⁴cs709/notes/LinearAlgebra.pdf

Example of linearly correlated features

- Example where ϕ doesn't have a full column rank,

$$\phi = \begin{bmatrix} x_1 & x_1^2 & x_1^2 & x_1^3 \\ x_2 & x_2^2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_n^2 & x_n^2 & x_n^3 \end{bmatrix} \quad (16)$$

- This is the simplest form of linear correlation of features, and it is not at all desirable.

Some questions

- Based on different inequalities between m and p , what are the cases where the least squares linear regression has (a) no solution (b) one solution and (c) multiple solutions.

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:

$$H_{\mathbf{v},\mathbf{p}} = \left\{ \mathbf{q} \mid (\mathbf{p} - \mathbf{q})^T \mathbf{v} = 0 \right\} \quad (17)$$

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}^* .

$$TH_{\mathbf{x}^*} = \left\{ \mathbf{p} \mid (\mathbf{p} - \mathbf{x}^*)^T \nabla f(\mathbf{x}^*) = 0 \right\} \quad (18)$$

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. $-\nabla F(\mathbf{a})$.

Therefore,

$$\Delta \mathbf{w}^{(k)} = -\nabla \varepsilon(\mathbf{w}^{(k)}) \quad \text{from equation (10)}$$

Hence,

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + 2\mathbf{t}^{(k)}(\phi^T \mathbf{y} - \phi^T \phi \mathbf{w}^{(k)}) \quad (19)$$

Gradient Descent Algorithm

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

- $\Delta \mathbf{w}^k = -\nabla \varepsilon(\mathbf{w}^{(k)})$
- Choose a step size $t^{(k)} > 0$ using exact or backtracking ray search.
- Obtain $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + t^{(k)} \Delta \mathbf{w}^{(k)}$.
- 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 \mathbf{x} should move along that direction.
- In general,

$$t^{(k)} = \underset{t}{\operatorname{argmin}} f\left(\mathbf{w}^{(k+1)}\right) \quad (20)$$

- Thus,

$$t^{(k)} = \underset{t}{\operatorname{argmin}} \left(\mathbf{w}^{(k)} + 2t \left(\phi^T \mathbf{y} - \phi^T \phi \mathbf{w}^{(k)} \right) \right) \quad (21)$$

Example of Gradient Descent Algorithm

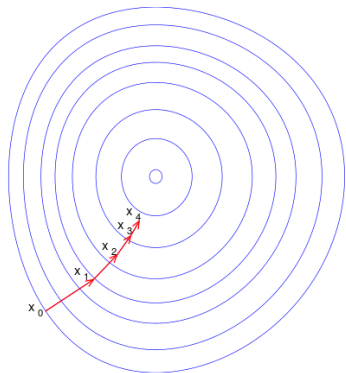


Figure: A red arrow originating at a point shows the direction of the negative gradient at that point. Note that the (negative) gradient at a point is orthogonal to the level curve going through that point. We see that gradient descent leads us to the bottom of the bowl, that is, to the point where the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