Introduction to Machine Learning - CS725

Instructor: Prof. Ganesh Ramakrishnan

Lecture 08 - Optimization Foundations Applied to Regression Formulations

Building on questions on Least Squares Linear Regression

1. Is there a probabilistic interpretation?
   - Gaussian Error, Maximum Likelihood Estimate
2. Addressing overfitting
   - Bayesian and Maximum Aposteriori Estimates, Regularization, Support Vector Regression
3. How to minimize the resultant and more complex error functions?
   - Level Curves and Surfaces, Gradient Vector, Directional Derivative, Gradient Descent Algorithm, Convexity, Necessary and Sufficient Conditions for Optimality

Lagrange fn, KKT conditions, Dual formu-lation

SVR objective

- 1-norm Error, and L_2 regularized:
  - $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$
    s.t. $\forall i$,
    $$\left.\begin{aligned} y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b &\leq \epsilon + \xi_i, \\ b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i &\leq \epsilon + \xi_i^*, \end{aligned}\right\} \text{4m constraints}$$
    $\xi_i, \xi_i^* \geq 0$
- 2-norm Error, and L_2 regularized:
  - $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i^2 + \xi_i^{*2})$
    s.t. $\forall i$,
    $$\left.\begin{aligned} y_i - \mathbf{w}^\top\phi(x_i) - b &\leq \epsilon + \xi_i, \\ b + \mathbf{w}^\top\phi(x_i) - y_i &\leq \epsilon + \xi_i^* \end{aligned}\right\} \text{2m constraints}$$
  - Here, the constraints $\xi_i, \xi_i^* \geq 0$ are not necessary

Need for Optimization so far (Equivalences we need to show)

You just cannot avoid optimization in ML!

- **Unconstrained (Penalized) Optimization:**

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \ ||\Phi\mathbf{w} - \mathbf{y}||_2^2 + \underbrace{\Omega(\mathbf{w})}_{\lambda}$$

- **Constrained Optimization 1:**

$$\mathbf{w}_{Reg} = \arg\min_{\mathbf{w}} \ ||\Phi\mathbf{w} - \mathbf{y}||_2^2$$

$$\textit{such that } \Omega(\mathbf{w}) \leq \theta$$

- **Constrained Optimization 2 ($t = 1$ or 2):**

$$\arg\min_{\mathbf{w}, b, \xi_i, \xi_i^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i (\xi_i^t + \xi_i^{*t})$$

s.t. $\forall i,\ y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i;\ b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$

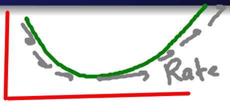
- **Equivalence**: λ (Penalized) $\equiv$ θ (Constrained)
- **Duality**: Dual of Support Vector Regression

Solving Unconstrained Minimization Problem

Derivative increases!

High school calculus

$\frac{df}{dx} = 0$



Rate of change = 0 at min

- Intuitively: Minimize by setting derivative (gradient) to 0 and hoping to find **closed form** solution.

$\frac{d^2f}{dx^2} \geq 0$ (Necessary) $\frac{d^2f}{dx^2} > 0$ (sufficient)

- When is such a solution a global minimum?
- For most optimization problems, finding closed form solutions 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 iterative methods?

$\frac{df}{dx}$ increases $\Rightarrow \frac{d^2f}{dx^2} \geq 0$

$\frac{df}{dx}$ strictly increases $\Rightarrow \frac{d^2f}{dx^2} > 0$

$\{x \mid f(x) \leq c\} \rightarrow$ Sublevel set
$\rightarrow$ set of pts encapsulated by level curve

- 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 its argument x.
- Formally we can define a level curve as :

$$L_c(\mathbf{f}) = \left\{ \mathbf{x} | \mathbf{f}(\mathbf{x}) = \mathbf{c} \right\} \tag{1}$$

where c is a constant.

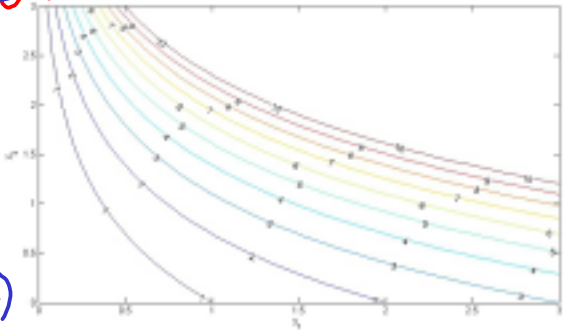
Foundations: Level curves and surfaces

- Example of different level curves for a single function

Q: What does gap between consecutive level curves indicate?

Ans: Larger the gap, smaller is rate of change!

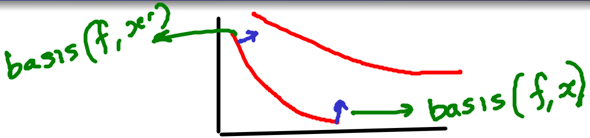
Region of fast rate of change (quantify using Directional derivative)



Region of slow rate of change

Figure 1: 10 level curves for the function $\mathbf{f}(\mathbf{x_1}, \mathbf{x_2}) = \mathbf{x_1 e^{x_2}}$ (Figure 4.12 from https://www.cse.iitb.ac.in/~CS725/notes/classNotes/BasicsOfConvexOptimization.pdf)

Foundations: Directional Derivatives



- Directional derivative: Rate at which the function changes at a given point $\mathbf{x}$ in a given direction $\mathbf{v}$
- 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 :

measured at x   measured along v

$$D_{\mathbf{v}}(f, \mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{v}) - \mathbf{f}(\mathbf{x})}{h} \tag{2}$$

$$s.t. \ ||\mathbf{v}||_2 = \mathbf{1} \tag{3}$$

(Claim)

Any $D_v(f, x) = \sum_i v_i \underbrace{[\text{basis}(f, x)]_i}_{\text{Gradient}} = v^T \underbrace{\nabla f(x)}_{\text{Gradient}}$

Foundations: Gradient Vector

- The **g**radient vector of a function f at a point $\mathbf{x}$ is defined as:

$$\nabla f_{\mathbf{x}^*} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ . \\ . \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \epsilon \mathbb{R}^n \qquad (4)$$

$$\max_{v} D_v(f, x) = \|\nabla f(x)\|$$

- 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.

$$\operatorname{argmax}_{v} D_v(f, x) = \frac{1}{\|\nabla f(x)\|} \nabla f(x)$$

Foundations: Gradient Vector

- The figure below illustrates the gradient vector for the same level curves

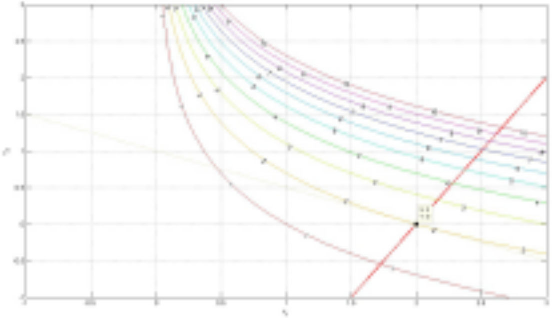
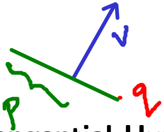


Figure 2: The level curves along with the gradient vector at (2, 0). Note that the gradient vector is perpenducular to the level curve $x_1 e^{x_2} = 2$ at (2, 0)

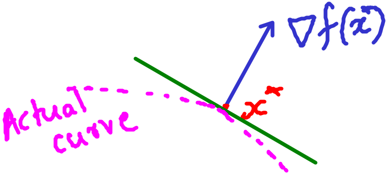
Hyperplanes

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



$$H_{\mathbf{v},\mathbf{q}} = \left\{ \mathbf{p} \mid (\mathbf{p} - \mathbf{q})^{\mathsf{T}} \mathbf{v} = \mathbf{0} \right\} \tag{5}$$

- **Tangential Hyperplane:** Plane orthogonal to the gradient vector at $\mathbf{x}^*$.



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 disjoint half-spaces.
- Technically, a hyperplane is a set of points whose direction *w.r.t.* a point $\mathbf{q}$ is orthogonal to a vector $\mathbf{v}$:

$$H_{\mathbf{v},\mathbf{q}} = \left\{ \mathbf{p} \mid (\mathbf{p} - \mathbf{q})^{\mathbf{T}} \mathbf{v} = \mathbf{0} \right\} \tag{5}$$

- **Tangential Hyperplane:** Plane orthogonal to the gradient vector at $\mathbf{x}^*$.

$$TH_{\mathbf{x}^*} = \left\{ \mathbf{p} \mid (\mathbf{p} - \mathbf{x}^*)^{\mathbf{T}} \nabla \mathbf{f}(\mathbf{x}^*) = \mathbf{0} \right\} \tag{6}$$

(Ridge Regression)

We recall that the problem was to find **w** such that

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\Phi\mathbf{w} - \mathbf{y}\|^2 + \lambda||\mathbf{w}||^2 \tag{7}$$

$$= \arg\min_{\mathbf{w}} (\underbrace{\mathbf{w}^T\Phi^T\Phi\mathbf{w} - 2\mathbf{w}^T\phi\mathbf{y} - \mathbf{y}^T\mathbf{y}}_{L_R(w,\mathcal{D})} + \lambda||\mathbf{w}||^2) \tag{8}$$

Foundations: Gradient Vector

- Magnitude (euclidean norm) of gradient vector at any point indicates maximum value of directional derivative at that point
- Thus, at the point of minimum of a differentiable minimization objective (such as least squares for regression), ...

If w^* is soln to Ridge Regression, it must minimize $\underbrace{\text{Ridge Regression}}_{L_R(w^*;D)}$ Loss $\Rightarrow \nabla_w L_R(w^*;D) = 0$

- *If $\nabla f(\mathbf{w}^*)$ is defined & $\mathbf{w}^*$ is local minimum/maximum, then* $\nabla f(\mathbf{w}^*) = 0$ (A necessary condition) (Cite : Theorem 60) of CS725/notes/classNotes/BasicsOfConvexOptimization.pdf
- Given that

Think of $\phi \in \mathbb{R}$ & $w \in \mathbb{R} \Rightarrow f(w) = \phi^2 w^2 - 2w\phi y - y^2 + \lambda w^2$

$f'(w) = 2\phi^2 w - 2\phi y + 2\lambda w$

$$f(\mathbf{w}) = (\mathbf{w}^T \Phi^T \Phi \mathbf{w} - 2\mathbf{w}^T \Phi^T \mathbf{y} - \mathbf{y}^T \mathbf{y} + \lambda ||\mathbf{w}||^2)$$

$$\Longrightarrow \ldots\ldots\ldots\ldots$$

$\nabla f(w) = 2\phi^T \phi w - 2\phi^T y + 2\lambda w$

Tricks: $\nabla_x (x^T A x) = 2Ax$

- We would have

Note

$y^T A x \in \mathbb{R}$

$\Rightarrow (y^T A x)^T = (y^T A x)$

$\nabla f(w^*) = 0 \Rightarrow$

$$\Longrightarrow \ldots\ldots\ldots\ldots$$

$w^* = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T y$

$$\Longrightarrow \ldots\ldots\ldots\ldots$$

$\nabla_x (x^T A) = A$

Thus: $\nabla_x (y^T A x) = \nabla_x (x^T A^T y) = A^T y$

Foundations: Necessary condition 1

- *If* $\nabla f(\mathbf{w}^*)$ *is defined &* $\mathbf{w}^*$ *is local minimum/maximum, then* $\nabla f(\mathbf{w}^*) = 0$ (A necessary condition) (Cite : Theorem 60) `CS725/notes/classNotes/BasicsOfConvexOptimization.pdf`
- Given that

$$f(\mathbf{w}) = (\mathbf{w}^T \Phi^T \Phi \mathbf{w} - 2\mathbf{w}^T \Phi^T \mathbf{y} - \mathbf{y}^T \mathbf{y} + \lambda ||\mathbf{w}||^2) \tag{9}$$

$$\implies \nabla f(\mathbf{w}) = 2\Phi^T \Phi \mathbf{w} - 2\Phi^T \mathbf{y} + 2\lambda \mathbf{w} \tag{10}$$

- We would have

$$\nabla f(\mathbf{w}^*) = 0 \tag{11}$$

$$\implies 2(\Phi^T \Phi + \lambda I)\mathbf{w}^* - 2\Phi^T \mathbf{y} = 0 \tag{12}$$

$$\implies \mathbf{w}^* = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T \mathbf{y} \tag{13}$$

Foundations: Necessary Condition 2

$\nabla^2 f(w^*) \succeq 0$ if $D_v f(x)$ increases in all directions v around w^*

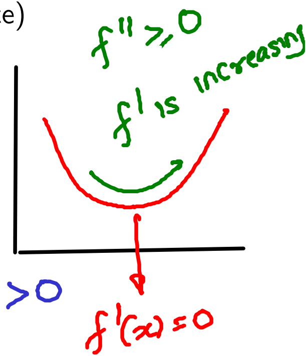
- *Is* $\nabla^2 f(\mathbf{w}^*)$ *positive definite ?*
  i.e. $\forall \mathbf{x} \neq 0$, *is* $\mathbf{x}^T \nabla f(\mathbf{w}^*)\mathbf{x} > 0$? (A sufficient condition for local minimum)
  (Note : Any positive definite matrix is also positive semi-definite)
  (Cite : Section 3.12 & 3.12.1)[1]

Trick: $\nabla^2 f(x^T A x) = 2A$

Hessian Matrix

$$\nabla^2 f(w) = \begin{bmatrix} \frac{d^2 f}{dw_1} & & \frac{\partial^2 f}{\partial w_i \partial w_j} \\ & \frac{d^2 f}{dw_2^2} \ddots & \\ \vdots & & \frac{\partial^2 f}{\partial w_p^2} \end{bmatrix} \Longrightarrow \nabla^2 f(w^*) = 2\Phi^T\Phi + 2\lambda I$$

$\Phi^T\Phi$ is p.s.d and with $\lambda > 0$, $(\Phi^T\Phi + \lambda I) \succ 0$



- And if Φ **has full column rank**, ...

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

[1] CS725/notes/classNotes/LinearAlgebra.pdf

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

$$
\begin{aligned}
\nabla^2 f(\mathbf{w}^*) &= 2\Phi^T\Phi + 2\lambda I && (14)\\
\implies \mathbf{x}^T \nabla^2 f(\mathbf{w}^*)\mathbf{x} &= 2\mathbf{x}^T(\Phi^T\Phi + \lambda I)\mathbf{x} && (15)\\
&= 2\left((\Phi + \sqrt{\lambda} I)\mathbf{x}\right)^T \Phi\mathbf{x} && (16)\\
&= 2\left\|(\Phi + \sqrt{\lambda} I)\mathbf{x}\right\|^2 \geq 0 && (17)
\end{aligned}
$$

- And with $\lambda = 0$, if Φ **has full column rank** ,

$$\Phi\mathbf{x} = 0 \quad \textit{iff} \quad \mathbf{x} = 0 \tag{18}$$

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

[2] `CS725/notes/classNotes/LinearAlgebra.pdf`

Example of linearly correlated features

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

$$\Phi = \begin{bmatrix} x_1 & x_1^2 & x_1^2 & x_1^3 \\ x_2 & x_2^2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_n^2 & x_n^2 & x_n^3 \end{bmatrix} \tag{19}$$

→ Not full column rank

$\Phi^T\Phi \geq 0$

Yet

- This is the simplest form of linear correlation of features, and it is not at all desirable.
- Effect of a nonzero λ with such Φ is that

$\Phi^T\Phi + \lambda I > 0$

Example of linearly correlated features

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

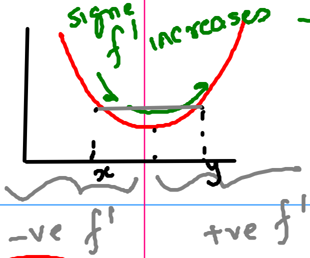
$$\Phi = \begin{bmatrix} x_1 & x_1^2 & x_1^2 & x_1^3 \\ x_2 & x_2^2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & x_n^2 & x_n^2 & x_n^3 \end{bmatrix} \tag{19}$$

- This is the simplest form of linear correlation of features, and it is not at all desirable.
- Effect of a nonzero λ with such Φ is that it tends to make the Hessian more positive definite

Do Closed-form solutions Always Exist?

- Linear regression and Ridge regression both have closed-form solutions
  - For linear regression,
$$w^* = (\Phi^\top \Phi)^{-1} \Phi^\top y$$
  - For ridge regression,
$$w^* = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top \mathbf{y}$$
    (for linear regression, $\lambda = 0$)
- What about optimizing the formulations (constrained/penalized) of Lasso (L_1 norm)? And support-based penalty (L_0 norm)?: Also requires tools of Optimization/duality

$\nabla\left(w^T\Phi^T\Phi w + 2w^T\Phi y + y^T y + \lambda\|w\|_1\right)$ does not exist!



$\nabla^2 f > 0 \longleftrightarrow f$ is strictly convex

$\nabla^2 f \geq 0 \longleftrightarrow f$ is convex

Cup shaped

fn is always below the gray line segment

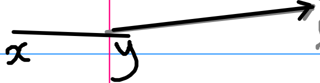
If $\nabla^2 f(x) \geq 0 \;\; \forall \, x \in D$ then f is strictly convex in D

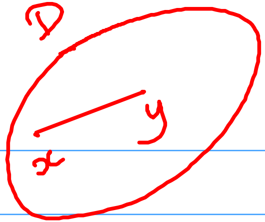
|||

If $x, y \in D$, $\theta x + (1-\theta) y \in D$

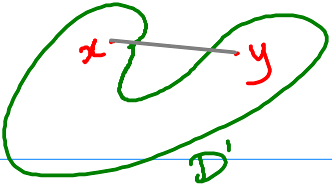
D is also convex set

If $f(\underbrace{\theta x + (1-\theta) y}_{\text{for } \theta \in [0,1] \text{ is a pt on line segment } [x,y]}) \leq \theta f(x) + (1-\theta) f(y)$



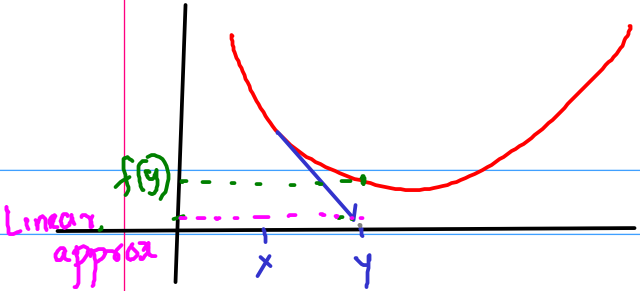


Convex D



Line segment $[x,y]$ is NOT completely contained in D'

$\Rightarrow$ D' is NOT convex



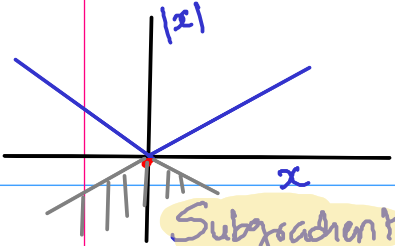
Observation for convex fn:

$$\underbrace{\left(\begin{array}{l}\text{Linear approx at } x\\ \text{evaluated at } y\end{array}\right)}_{f(x)\,+\,\nabla^T f(x)(y-x)} \leq f(y)$$

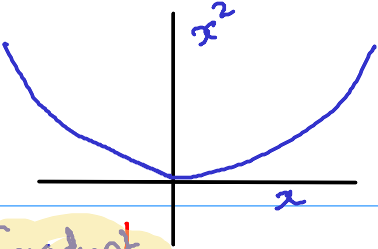
strict convexity

convex

$\nabla^T f(x)(y-x) \rightarrow$ Accounts for slope



Subgradient



Gradient

Think of a vector that behaves like gradient vec:

③ ∇ is direction that gives lower bnd linear approx

① ∇ is direction of max rate of change

② $\perp$ to tangent hyperplane

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 $\mathbf{w^{(k+1)}} = \mathbf{w^{(k)}} + \mathbf{t^{(k)}}\boldsymbol{\Delta}\mathbf{w^{(k)}}$.
- Set $k = k + 1$. **until** stopping criterion
  (such as $\|\nabla\varepsilon(\mathbf{w}^{(k+1)})\| \leq \epsilon$) is satisfied