Foundations of Machine Learning
AI2000 and AI5000

FoML-21
Logistic Regression - Newton Raphson optimization

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad
July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

So far in FoML

- Intro to ML and Probability refresher
- MLE, MAP, and fully Bayesian treatment
- Supervised learning
    a. Linear Regression with basis functions (regularization, model selection)
    b. Bias-Variance Decomposition (Bayesian Regression)
    c. Decision Theory - three broad classification strategies
        - Probabilistic Generative Models - Continuous & discrete data
        - (Linear) Discriminant Functions - least squares solution, Perceptron
        - Probabilistic Discriminative Models - Logistic Regression

Logistic Regression - IRLS

Loss function approximation

- Taylor series expansion (univariate case)

$$E(w+\Delta w) = E(w) + \Delta w \cdot E'(w) + \frac{1}{2}\Delta w^2 E''(w) + \cdots\cdots\cdots$$

Loss function approximation

- Taylor series expansion (multivariate case)

$$E(\mathbf{w}+\Delta\mathbf{w}) = E(\underline{w}) + \underline{\Delta w}^{T} \cdot \underline{\nabla E} + \frac{1}{2}\, \underline{\Delta w}^{T} \cdot \underline{H} \cdot \underline{\Delta w} + \ldots\ldots\ldots$$

$$H_{ij} = \frac{\partial^2 E(\underline{w})}{\partial w_i \, \partial w_j}$$

Hessian of E at w

Linear (first-order) approximation

$$E(\mathbf{w}+\Delta\mathbf{w}) \approx E(\mathbf{w}) + \Delta\mathbf{w}^T \nabla E(\mathbf{w})$$

for $\downarrow < E(w)$

$\Delta\mathbf{w}^T \nabla E(\mathbf{w}) < 0$

$\Rightarrow \Delta w = -\eta \nabla E(w)$ { i.e., take a small step in the opposite direction of gradient

Quadratic (second-order) approximation

$$E(\mathbf{w}+\Delta\mathbf{w}) \approx E(\mathbf{w}) + \Delta\mathbf{w}^T \nabla E(\mathbf{w}) + \frac{1}{2}\Delta\mathbf{w}^T \nabla^2 E(\mathbf{w}) \Delta\mathbf{w}$$

what is the best Δw for minimizing $E(w)$

Hessian

$$\Rightarrow \frac{\partial E(w+\Delta w)}{\partial \Delta w} = 0 \quad \Rightarrow \quad \nabla E(w) + \nabla^2 E(w) \cdot \Delta w = 0$$

$$\Delta w = -\left[\nabla^2 E(w)\right]^{-1} \nabla E(w)$$

$H \times H \quad H \times 1$

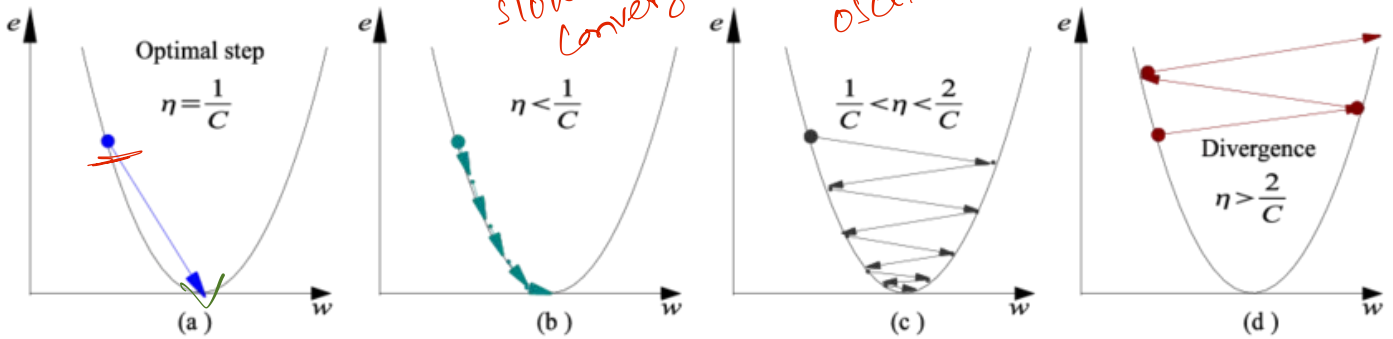
$\Leftarrow$ Best Δw to minimize $E(w)$

$$w^{t+1} = w^t - H^{-1} \nabla E(w)$$

Convergence for Quadratic functions

$E(w) = \frac{1}{2}cw^2 + bw + a$

$E''(w) = H = c$



quadratic loss function is Convex

$\Rightarrow$ 2nd order approximation will be exact for such functions

$\Rightarrow$ single step convergence (irrespective of w^0)

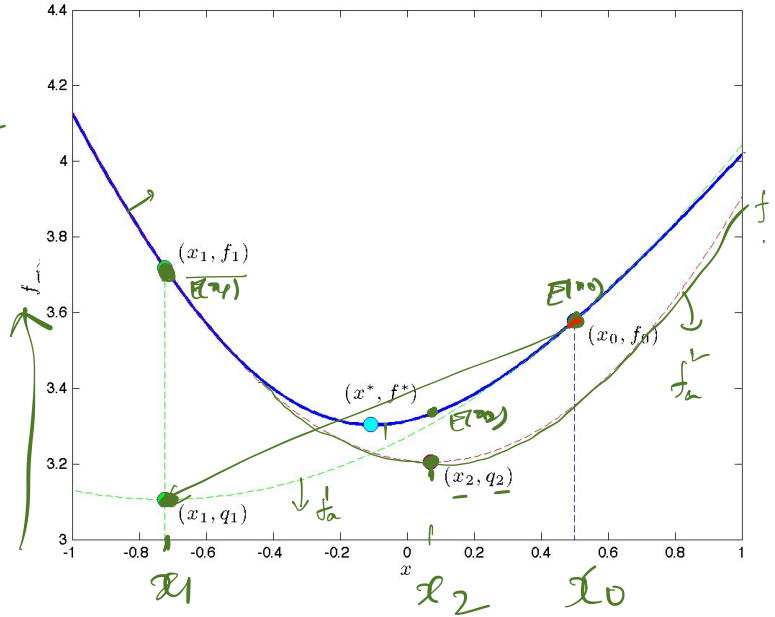
$$w^* = w^0 - H^{-1} \nabla E(w)$$

optimizing

Generic Convex functions

with Second-order approx.

example for generic (degree > 2) but convex loss function



→ x_0 = initial guess

$x_1 = x_0 - \frac{1}{f''(x_0)} f'(x_0)$

$x_2 = x_1 - \frac{1}{f''(x_1)} f'(x_1)$

Newton-Raphson Iterative optimization

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \mathbf{H}^{-1}\nabla E(\mathbf{w}^t)$$

$$\nabla E(\mathbf{w})^T = \sum_{n=1}^{N}(y_n - t_n)\phi_n = \Phi^T[\underline{y} - \underline{t}]$$

$$\phi_n = \underline{\phi}(x_n)$$

$$\Phi = \begin{bmatrix} -\phi(x_1)- \\ -\phi(x_2)- \\ \vdots \\ -\phi(x_N)- \end{bmatrix}_{N\times M} \begin{matrix} \times(y_1 - t_1) \\ \\ \\ \times(y_N - t_N) \end{matrix} \quad \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1} \quad \underline{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$

Newton-Raphson Iterative optimization

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \mathbf{H}^{-1}\nabla E(\mathbf{w}^t)$$

$$\mathbf{H}_{ij} = \frac{\partial^2 E(\mathbf{w}^t)}{\partial \mathbf{w}_i \partial \mathbf{w}_j} = \frac{\partial}{\partial \mathbf{w}_i}\sum_{n=1}^{N}(y_n - t_n)\phi_j(\mathbf{x}_n) = \sum_{n=1}^{N} y_n(1-y_n)\phi_i(x_n)\phi_j(x_n)$$

$$H = \Phi^T R \Phi$$

$M \times M \quad M \times N \quad N \times N \quad N \times M$

$$R_{nn} = y_n(1-y_n)$$

भारतीय सांकेतिक विज्ञान संस्थ हैदराबाद
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

Newton-Raphson Iterative optimization

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \mathbf{H}^{-1}\nabla E(\mathbf{w}^t)$$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - (\Phi^T R\Phi)^{-1}\Phi^T(\mathbf{y}-\mathbf{t}) = (\Phi^T R\Phi)^{-1}\Phi^T R\mathbf{z}$$

$$= (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t}$$

$R_{nn} = y_n(1-y_n)$ $\quad \sigma(w^T\phi_n) = y_n = P(C_1|x)$

Iterative Reweighted Least Squares

$$E(w) = \sum_{i=1}^{N} R_i (t_i - y_i)^2$$

Next
Neural Networks