LEC6 DSA5204

same epoch, more updates!

Re-cap:

- FCNN
- RNN
- CNN

$\downarrow$ $\boxed{\text{Model}}$

- GD
- SGD (computationally efficient)
- SGD with momentum

gradient

$\Downarrow$

"$\boxed{\text{external force}}$"

$\downarrow$ $\boxed{\text{Optimization Algorithm}}$

Future: ① How to improve performance?

② Outside super-vised learning scope.

---

① How to improve model performance?

a) Model Architecture

b) Training Method (Optimization)

c) Data (Augmentation)

1. Regularization

$\rightarrow$ low testing error

$\rightarrow$ probably high training error

$\Downarrow$

Inductive bias such that model can behave better with limited data

Previously: ERM Framework (Empirical Risk Minimization)

$$\hat{\theta}_{ERM} = \arg\min_{\theta} R_{emp}(\theta)$$

Regularization Framework

$$\hat{\theta}_{Reg} = \arg\min_{\theta} R_{emp}(\theta) + \alpha \underbrace{\Omega(\theta)}_{\text{regularization / penalty term}}$$

a) l_2-reg : $\Omega(\theta) = \frac{1}{2}\|\theta\|_2^2$

$\to$ Gaussian prior on w (statistics perspective)

Rmk: penalize more on large value than l_1-reg, But cannot guarantee sparsity!

Toy example: (Ridge (linear) regression)

$$\begin{cases} \Omega(w) = \frac{1}{2}\|w\|_2^2 \\ R_{emp}(w) = \frac{1}{2}\|Xw - y\|_2^2 \end{cases} \Longrightarrow$$

Property: have closed-form sol^n

$$\Rightarrow \hat{w}_{ridge} = (X^T X + \alpha I)^{-1} X^T y$$

Note: $\hat{w}_{LR} = (X^T X)^{-1} X^T y$

$X \in \mathbb{R}^{n \times m}$

$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$

$\Rightarrow X^T X = \sum_{i=1}^{n} x_i x_i^T$

$\downarrow$

Covariance matrix

We have: $\|\hat{w}_{LR}\| \geq \|\hat{w}_{ridge}\|$

Pf: consider $X^T X$ (symmetric matric)

$\downarrow$

eigenvalue decomp.

$$\begin{cases} \text{eigenvalue} & \lambda_1 \geq \cdots \geq \lambda_m \\ \text{eigenvector} & u_1, \ldots, u_m \end{cases}$$

Principal Component (direction)

then: $\hat{w}_{LR} = (X^T X)^{-1} X^T y$

$$= (X^T X)^{-1} \cdot \sum_{i=1}^{m} \beta_i u_i \quad \Rightarrow \|\hat{w}_{LR}\|_2^2 = \sum_{i=1}^{m} \frac{\beta_i^2}{(\lambda_i)^2}$$

$\boxed{\beta_i = \langle u_i, X^T y \rangle}$

$$= \sum_{i=1}^{m} \frac{\beta_i}{\lambda_i} u_i$$

$$\hat{w}_{ridge} = (X^T X + \alpha I)^{-1} X^T y$$

$$= (X^T X + \alpha I)^{-1} \sum_{i=1}^{m} \beta_i u_i \Rightarrow \|\hat{w}_{ridge}\|_2^2 = \sum_{i=1}^{m} \frac{\beta_i^2}{(\lambda_i + \alpha)^2}$$

$$= \sum_{i=1}^{m} \frac{\beta_i}{\lambda_i + \alpha} u_i$$

Note: if α is large, then $\frac{\beta_i}{\lambda_i + \alpha} \to 0$

$$\Rightarrow \hat{w}_{ridge} \to 0$$

Also, ℓ_2-regularization can be viewed as Weight Decay for non-linear case (we use GD)

$$\theta^{(k+1)} = \theta^{(k)} - \varepsilon \nabla_\theta R_{emp}(\theta^{(k)}) - \underbrace{\varepsilon \cdot \alpha\, \theta^{(k)}}_{\text{weight decay part}}$$

In application, Weight Decay $\Leftrightarrow$ ℓ_2-regularization

Statistics Perspective $\boxed{\text{Laplacian prior on } w}$

b) ℓ_1 - regularization (non-smooth but convex)

Toy example: (LASSO (linear) regression)

$$\begin{cases} \Omega(w) = \|w\|_1 \\ R_{emp}(w) = \frac{1}{2}\|Xw - y\|_2^2 \end{cases} \Rightarrow \boxed{\text{no closed-form sol}^{n}}$$

Example: ℓ_1 v.s. ℓ_2 regularization

$$R_{emp}(\theta) = \frac{1}{2}\sum_{i=1}^{m} \lambda_i (\theta_i - \theta_i^*)^2$$

$\boxed{\theta_i^* = \operatorname{argmin}_\theta R_{emp}(\theta)}$ → without reg

1. ℓ_2-norm: $\boxed{\tilde{R}_{emp}(\theta)} = \frac{1}{2}\sum_{i=1}^{m}\lambda_i(\theta_i - \theta_i^*)^2 + \frac{1}{2}\alpha\sum_{i=1}^{m}\theta_i^2$

convex + differentiable (smooth)

$$\frac{\partial \tilde{R}_{emp}}{\partial \theta_i} = \lambda_i(\theta_i - \theta_i^*) + \alpha\theta_i$$

(convexity)

$$\hat{\theta}^{\ell_2} \in \operatorname{argmin}_\theta \tilde{R}_{emp}(\theta) \Leftrightarrow \frac{\partial \tilde{R}_{emp}}{\partial \theta_i}(\hat{\theta}^{\ell_2}) = 0$$

$$\Rightarrow \hat{\theta}_i^{\ell_2} = \frac{\lambda_i}{\lambda_i + \alpha}\theta_i^*$$

2. ℓ_1-norm $\boxed{\tilde{R}_{emp}(\theta)} = \frac{1}{2}\sum_{i=1}^{m}\lambda_i(\theta_i - \theta_i^*)^2 + \alpha\sum_{i=1}^{m}|\theta_i|$

convex but non-smooth

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta} \tilde{R}_{emp}(\theta)$$

(highly non-trivial)

$$\Leftrightarrow \quad 0 \in \partial \tilde{R}_{emp}(\hat{\theta}^{\ell_1})$$

$$\Leftrightarrow \quad 0 \in \begin{pmatrix} \lambda_1(\hat{\theta}_1^{\ell_1} - \theta_1^*) \\ \vdots \\ \lambda_m(\hat{\theta}_m^{\ell_1} - \theta_m^*) \end{pmatrix} + \alpha \boxed{\partial \|\cdot\|_1(\hat{\theta}^{\ell_1})}$$

$$\partial \|\cdot\|_1(\hat{\theta}^{\ell_1}) = \begin{pmatrix} \partial \|\cdot\|_1(\hat{\theta}_1^{\ell_1}) \\ \vdots \\ \partial \|\cdot\|_1(\hat{\theta}_m^{\ell_1}) \end{pmatrix}$$

$$\boxed{\Leftrightarrow \quad 0_i \in (\hat{\theta}_i^{\ell_1} - \theta_i^*) + \frac{\alpha}{\lambda_i} \partial \|\cdot\|(\hat{\theta}_i^{\ell_1})}$$

$$\Leftrightarrow \quad \hat{\theta}_i^{(\ell)} = P_{\frac{\alpha}{\lambda_i}\|\cdot\|_1}(\theta_i^*)$$

$$= \begin{cases} \theta_i^* - \frac{\alpha}{\lambda_i}, & \theta_i^* \geq \frac{\alpha}{\lambda_i} \\ 0, & o/w. \\ \theta_i^* + \frac{\alpha}{\lambda_i}, & \theta_i^* \leq -\frac{\alpha}{\lambda_i} \end{cases}$$

$$\boxed{\begin{aligned} &u = P_f(x) = \arg\min_y f(y) + \tfrac{1}{2}\|y - x\|_2^2 \\ &\Leftrightarrow \quad 0 \in u - x + \partial f(u) \end{aligned}}$$

---

② Regularization on NN

→ we seldom regularize on bias term b

→ we may choose different strength of regularization for each layer

$$\boxed{\alpha_i \rightarrow i\text{-th layer}}$$

---

③ Early Stopping for NN → under certain assumption, it is equivalent to $\ell 2$-reg !

Early Stopping: $\boxed{\text{implicit regularization}}$

$\boxed{\text{require validation set}}$!!

Normally, we use validation set to monitor the time point to stop !

Variant 1 : record the optimal epoch number

stop criterion

↓

$\boxed{\text{retrain}}$

Variant 2 : continue training with full dataset after early stop

↓

record optimal loss function value (training)

stop criterion

[E.g.] Early Stop for Linear Reg

$$R_{emp}(\theta) = \frac{1}{2}\lambda(\theta-\theta^*)^2 \Rightarrow \nabla R_{emp}(\theta) = \lambda(\theta-\theta^*)$$

Consider GD : $\theta_{k+1} = \theta_k - \varepsilon\lambda(\theta_k - \theta^*)$

$$= (1-\varepsilon\lambda)\theta_k + \varepsilon\lambda\theta^*$$

$$\Rightarrow \theta_{k+1} = (1-\varepsilon\lambda)^{k+1}\theta_0 + [1-(1-\varepsilon\lambda)^{k+1}]\theta^*$$

→ Stop at iteration τ : $\hat{\theta} = \theta_\tau = \boxed{(1-\varepsilon\lambda)^\tau}\theta_0 + [1-(1-\varepsilon\lambda)^\tau]\theta^*$

(variant)

→ L2-regularization : $\tilde{R}(\theta) = \frac{1}{2}\lambda(\theta-\theta^*)^2 + \frac{1}{2}\alpha(\theta-\theta_0)^2$

$$\nabla\tilde{R}(\theta) = \lambda(\theta-\theta^*) + \alpha(\theta-\theta_0) = 0$$

$$\Rightarrow \tilde{\theta} = \boxed{\frac{\alpha}{\alpha+\lambda}}\theta_0 + \left(1-\frac{\alpha}{\alpha+\lambda}\right)\theta^*$$

Note $\hat{\theta} = \tilde{\theta} \Leftrightarrow \frac{\alpha}{\alpha+\lambda} = (1-\varepsilon\lambda)^\tau \Leftrightarrow \boxed{\alpha = \frac{\lambda(1-\varepsilon\lambda)^\tau}{1-(1-\varepsilon\lambda)^\tau}}$

$\boxed{\text{early stop}}$ ↔ equivalent ↔ $\boxed{\text{l2-regularization}}$

↓ $\boxed{\text{regularization strength}}$

under certain condition { LR model / manual regularization strength

④ Add Noise → maintain more structure of hidden space

↓ like VAE

$\boxed{\text{implicit regularization}}$

(inductive bias)

a) Adding Noise to Data Input

[E.g.] LR

$R(w) = \frac{1}{2}\|Xw - y\|_2^2$

$$\boxed{Z = \begin{pmatrix} z_1^T \\ \vdots \\ z_n^T \end{pmatrix}}$$

$X \to X + Z$ $\quad z_i \sim \mathcal{N}(0, \delta I)$ $\quad$ (Add Noise)

↓

$$R(w) = \underbrace{\frac{1}{2}\|Xw - y\|_2^2}_{\text{const}} + \underbrace{\frac{1}{2}\|Zw\|_2^2 + \underbrace{(Zw)^T(Xw - y)}_{\text{zero mean}}}_{\text{random}}$$

$$\Rightarrow \mathbb{E}_Z[R(w)] = \frac{1}{2}\|Xw - y\|_2^2 + \frac{1}{2}\mathbb{E}_Z[w^T Z^T Z w]$$

$$= \frac{1}{2}\|Xw - y\|_2^2 + \frac{1}{2} w^T \mathbb{E}_Z[Z^T Z] w$$

On average, adding noise is equivalent to ℓ_2-norm reg

$$= \frac{1}{2}\|Xw - y\|_2^2 + \frac{1}{2}\cdot n \, w^T \mathbb{E}_{z_i}[z_i z_i^T] w$$

$$= \frac{1}{2}\|Xw - y\|_2^2 + \frac{n}{2}\cdot \delta \, w^T w$$

$$= \frac{1}{2}\|Xw - y\|_2^2 + \frac{n\delta}{2}\|w\|_2^2$$

② Another approach ⇒ Label Smoothing

③ Adding Noise to Weight