HW2

Hardik Prajapati (2678294168)

1. CART for Regression :-

a)

$$J = \text{cost}\{(\underline{x}_i, y_i) \mid \underline{x}_i \in R_{m'}\} = \sum_{\underline{x}_i \in R_{m'}} (y_i - w_{m'})^2$$

($w_{m'}$: Prediction for input x_i)

$$\Rightarrow J(w_{m'}) = \sum_{\underline{x}_i \in R_{m'}} (y_i - w_{m'})^2$$

$$\Rightarrow \frac{\partial J}{\partial w_{m'}} = \sum_{\underline{x}_i \in R_{m'}} 2(y_i - w_{m'})(-1)$$

$$\Rightarrow 0 = \sum_{\underline{x}_i \in R_{m'}} (y_i - w_{m'})$$

$$\Rightarrow 0 = \sum_{\underline{x}_i \in R_{m'}} y_i - \sum_{\underline{x}_i \in R_{m'}} w_{m'}$$

$$\Rightarrow 0 = \sum_{\underline{x}_i \in R_{m'}} y_i - N_{R_{m'}} \cdot w_{m'}$$

$$\Rightarrow \boxed{w_{m'}^* = \frac{1}{N_{R_{m'}}} \sum_{\underline{x}_i \in R_{m'}} y_i}$$

b) Given Region $R_{m'}$, containing $N_{R_{m'}}$ data points

Given x_j (feature).

No. of $t_k \leq$ Unique data points in $N_{R_{m'}}$

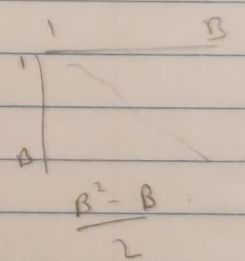
eg. $N_{R_{m'}} \in \{-10, 2, -5, 2, 4\}$ ← for the selected feature

Then $t_k \in \{-10, -5, 2, 4\}$

Iterate over each point such that we optimizes cost function.

2. $V_i \in$ i.i.d , mean μ_v & variance σ_v^2

$$S = \frac{1}{B} \sum_{i=1}^{B} V_i$$



a) correlation coefficient = ρ

$$\rho = \frac{E\{V_i V_j\} - \mu_v^2}{\sigma_v^2}, \quad \rho \geq 0$$

To show, variance of $S = \sigma_s^2 = \rho\sigma_v^2 + (1-\rho)\frac{\sigma_v^2}{B}$

$$Var(S) = Var\left(\frac{1}{B}\sum_{i=1}^{B} V_i\right)$$

$$= \frac{1}{B^2}\left\{\sum_{i=1}^{B} Var(V_i) + 2\sum_{i}\sum_{j<i} cov(X_i, X_j)\right\}$$

$$\Rightarrow Var(S) = \frac{1}{B^2}\left\{B\cdot\sigma_v^2 + 2\left(\frac{B^2-B}{2}\right)\rho\sigma_v\cdot\sigma_v\right\}$$

$$= \frac{1}{B^2}\left\{B\cdot\sigma_v^2 + B^2\rho\sigma_v^2 - B\rho\sigma_v^2\right\}$$

$$= \frac{1}{B^2}\left\{B\sigma_v^2\{1-\rho\} + B^2\rho\sigma_v^2\right\}$$

$$= \frac{\sigma_v^2}{B}(1-\rho) + \rho\sigma_v^2$$

2 b) $$\rho = \frac{Cov(V_i, V_j)}{(\sigma_v)\cdot(\sigma_v)}$$

By Cauchy-Schwartz inequality,

$$|Cov(V_i, V_j)|^2 \leq Var(V_i)Var(V_j)$$

$$\therefore |Cov(V_i, V_j)| \leq \sqrt{\sigma_v^2\sigma_v^2}$$

$$\therefore |Cov(V_i, V_j)| \leq \sigma_v^2$$

$$\Rightarrow \rho = \frac{Cov(V_i, V_j)}{\sqrt{Var(V_i)Var(V_j)}} \leq \frac{\sqrt{Var(V_i)Var(V_j)}}{\sqrt{Var(V_i)Var(V_j)}} = 1$$

$$\Rightarrow \rho \leq 1$$

c) $\sigma_s^2 = \rho \cdot \sigma_v^2 + (1-\rho)\frac{\sigma_v^2}{B}$

Given, σ_v^2 & B.

We know that $0 \leq \rho \leq 1$

↳ Substituting, $\rho = 0$,

$\Rightarrow \sigma_s^2 = 0 + \frac{\sigma_v^2}{B}$

$\Rightarrow \boxed{\sigma_s^2 = \frac{\sigma_v^2}{B}}$ ← Smallest variance of s.

↳ Substituting $\rho = 1$

$\Rightarrow \sigma_s^2 = \sigma_v^2 + 0$

$\Rightarrow \boxed{\sigma_s^2 = \sigma_v^2}$ ← Highest variance of s.

↳ Properties of Decision Trees in Random Forest:-

↦ Taking average over a set of Trees, we are able to reduce variance.

↦ Lower pairwise-correlation between trees, gives more reduction in variance.

4. Forward Stagewise Additive Modeling

a)

Eq 16.34 :- $f_m(x) = f_{m-1}(x) + \beta_m \phi(x; \gamma_m)$

Iteration 1:

$$f_1(x) = f_0(x) + \beta_1 \phi(x; \gamma_1)$$

Now,

$$(\beta_2, \gamma_2) = \underset{\beta, \gamma}{\arg\min} \sum_{i=1}^{N} L\left(y_i, f_1(x_i) + \beta \phi(x_i; \gamma)\right)$$

Iteration 2:

$$f_2(x) = f_1(x) + \beta_2 \phi(x; \gamma_2)$$

$$\Rightarrow f_2(x) = f_0(x) + \beta_1 \phi(x; \gamma_1) + \beta_2 \phi(x; \gamma_2)$$

Iteration k^{th}:

$$f_k(x) = f_0(x) + \sum_{i=1}^{k} \beta_i \phi(x; \gamma_i)$$

↑

This is of the form

Eq 16.3 :- $f(x) = w_0 + \sum_{m=1}^{M} w_m \phi_m(x)$

$\Rightarrow$ Forward Stagewise Additive modeling is an ABM

i) $w_0 = f_0(x)$ iii) $\phi_m(\underline{x}) = \phi(\underline{x}, \gamma_i)$

ii) $w_m = \beta_i$

4. b) Eq. 16.35 :- $f_m(x) = f_{m-1}(x) + \nu \beta_m \phi(x; \gamma_m)$

Iteration 1 :-

$$f_1(x) = f_0(x) + \nu \beta_1 \phi(x, \gamma_1)$$

where $(\beta_1, \gamma_1) = \underset{\beta, \gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, f_0(x_i) + \nu \beta \phi(x_i, \gamma))$

Now, Iteration 2 :-

$$(\beta_2, \gamma_2) = \underset{\beta, \gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, f_1(x_i) + \nu \beta \phi(x_i, \gamma))$$

$$f_2(x) = f_1(x) + \nu \beta_2 \phi(x, \gamma_2)$$

$$\Rightarrow f_2(x) = f_0(x) + \nu \beta_1 \phi(x, \gamma_1) + \nu \beta_2 \phi(x, \gamma_2)$$

Iteration k^{th} :-

$$\Rightarrow f_k(x) = f_0(x) + \sum_{j=1}^{k} \nu \beta_j \phi(x, \gamma_j)$$

This is of the form

Eq. 16.3 :- $f(x) = \omega_0 + \sum_{m=1}^{M} \omega_m \phi_m(x)$

$\Rightarrow$ Stagewise additive modelling with shrinkage is also ADM.

i) $\omega_0 = f_0(x)$ iii) $\phi_m(x) = \phi(x, \gamma_j)$

ii) $\omega_m = \nu \beta_j$

5. Adaboost. M1 , For Binary Classification with exponential Loss.

1) $w_i = 1/N$

2) for $m = 1 : M$ do

3) Fit a classifier $\phi_m(x)$ to the training set using weights w_i;

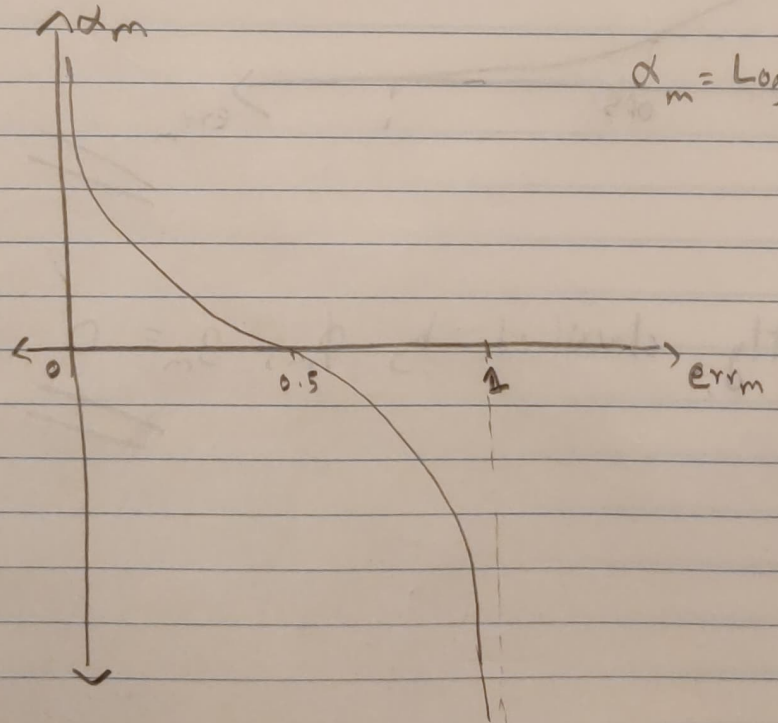
4) Compute $err_m = \dfrac{\sum_{i=1}^{N} w_{i,m} \, \mathbb{I}(\tilde{y}_i \neq \phi_m(x_i))}{\sum_{i=1}^{N} w_{i,m}}$;

5) Compute $\alpha_m = \text{Log}\left[(1 - err_m)/err_m \right]$;

6) Set $w_i \leftarrow w_i \exp\left[\alpha_m \, \mathbb{I}(\tilde{y}_i \neq \phi_m(x_i))\right]$;

7) Return $f(x) = \text{sgn}\left[\sum_{m=1}^{M} \alpha_m \phi_m(x) \right]$;

a) $\alpha_m = \text{Log}\left[\dfrac{(1 - err_m)}{err_m} \right]$, $0 < err_m < 1$



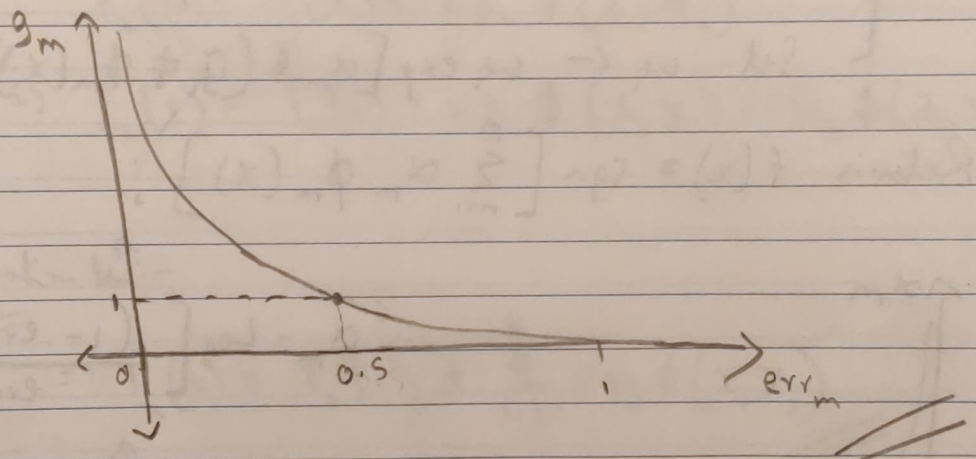
5. b

$$w_{i,m+1} = g_m w_{i,m}$$

$$g_m = \exp\left[\alpha_m \mathbb{I}\left(\tilde{y}_i \neq \phi_m(x_i)\right)\right]$$

$$= \exp\left[\log\left[\frac{(1-err_m)}{err_m}\right] \mathbb{I}\left(\tilde{y}_i \neq \phi_m(x_i)\right)\right]$$

$$\Rightarrow g_m = \begin{cases} \dfrac{1-err_m}{err_m} & , \text{if } \tilde{y}_i \neq \phi_m(x_i) \leftarrow \text{Wrongly Classified} \\ 0 & , \text{if } \tilde{y}_i = \phi_m(x_i) \leftarrow \text{Rightly Classified} \end{cases}$$



↳ For correctly classified by ϕ_m, $g_m = 0$

5. c

$$\alpha_m = \log\left(\frac{1-err_m}{err_m}\right)$$

↳ If $err_m < 0.5$

Then

→ $\alpha_m \in (0, \infty)$

Now,

→ $\beta_m = \frac{1}{2}\alpha_m \Rightarrow \beta_m \in (0, \infty)$

→ $g_m \in (1, \infty)$