Emily
Barnes
10/13/19

Homework 3

Question 1

CART modeling for regression

old tree, T_{old} has $|T_{old}| = M$ terminal nodes/buckets
↳ each bucket $m = 1, \dots, M$
let 1. N_m denote num of observations in bucket m
2. $Q_m(T_{old})$ denotes val of impurity function at bucket m
3. R_m denote the region in the feature space corresponding to bucket

N = overall # of observations

$$Q_m(T_{old}) = \frac{1}{N_m} \sum_{i:x_i \in R_m} (y_i - \hat{y}_m)^2$$

$\hat{y}_m = \frac{1}{N_m} \sum_{i:x_i \in R_m} y_i$ = mean response in bucket m

total impurity cost of the tree T_{old}:

$$C_{imp}(T_{old}) = \sum_{m=1}^{M} N_m Q_m(T_{old})$$

↳ consider split at final bucket M
↳ new tree T_{new} $|T_{new}| = M+1$
$\tilde{N}_m$ denotes new # observations in bucket m
$\tilde{Q}_m(T_{new})$ = impurity function at bucket m
$\tilde{R}_m$ denotes region in the feature space corresponding to bucket m

Total impurity cost of tree T_{new}

$$C_{imp}(T_{new}) = \sum_{m=1}^{M+1} \tilde{N}_m Q_m(T_{new})$$

1. a) let $\Delta = C_{imp}(T_{old}) - C_{imp}(T_{new})$ - absolute decrease in total impurity
↳ only depend on R_m

$$\Delta = C_{imp}(T_{old}) - C_{imp}(T_{new})$$

$$\Delta = \sum_{m=1}^{M} N_m Q_m (T_{old}) - \sum_{m=1}^{M} \tilde{N}_m \tilde{Q}_m (T_{new})$$

↳ buckets $m=1$ to $m=M-1$ are identical; only bucket M was altered

$$\Delta = \left(\left(\sum_{m=1}^{M-1} N_m Q_m(T_{old})\right) + N_M Q_M (T_{old})\right) - \left(\sum_{m=1}^{M-1} \tilde{N}_m \tilde{Q}_m(T_{new}) + \tilde{N}_M \tilde{Q}_M(T_{new}) + \tilde{N}_{M+1} \tilde{Q}_{M+1}(T_{new})\right)$$

→ for $m=1 \to m=M-1$, nodes are unchanged, ∴ $\hat{y}_m$ = mean response in each bucket is the same, ∴ $Q_m(T_{old}) = \tilde{Q}_m(T_{new})$ for all $m \leq M-1$, ∴ Δ function reduces to the following

$$\Delta = N_M Q_M (T_{old}) - \left(\tilde{N}_M \tilde{Q}_M (T_{new}) + \tilde{N}_{M+1} \tilde{Q}_{M+1}(T_{new})\right)$$

$$\Delta = N_M \left[\frac{1}{N_M} \sum_{i: x_i \in R_m} (y_i - \hat{y}_M)^2\right] - \left(\tilde{N}_M \frac{1}{\tilde{N}_M} \sum_{i: x_i \in R_m} (y_i - y_m)^2 + \tilde{N}_{M+1} \frac{1}{\tilde{N}_{M+1}} (y_i - y_{M+1})\right)$$

$$\Delta = \sum_{i=1}^{k} (y_i - \hat{y}_M)^2 - \sum_{i=1}^{j} (y_i - \hat{\hat{y}}_M)^2 - \sum_{i=j}^{k} (y_i - \hat{\hat{y}}_{M+1})$$

where the indicies of $1 \to k$ indicate all of the values in the original bucket M, with $i=1 \to j$ indicating the subset of these values which are split into M_{new} and $i=j \to k$ indicating the subset of these values which are split into the $M+1$ bucket

$$\Delta = SSE_{old} - SSE_{new}$$

b) $\Delta = \sum_{i=1}^{k} (y_i - \hat{y}_m)^2 - \sum_{i=1}^{j} (y_i - \tilde{\hat{y}}_m)^2 - \sum_{i=j}^{k} (y_i - \tilde{\hat{y}}_{m+1})^2$

$$\Delta = \sum_{i=1}^{j} (y_i - \hat{y}_m)^2 + \sum_{i=j}^{k} (y_i - \hat{y}_m) - \sum_{i=1}^{j} (y_i - \tilde{\hat{y}}_{\tilde{m}}) - \sum_{i=j}^{k} (y_i - \tilde{\hat{y}}_{m+1})$$

$$\Delta = \left[\sum_{i=1}^{j} (y_i - \hat{y}_m)^2 - \sum_{i=1}^{j} (y_i - \tilde{\hat{y}}_m)^2\right] + \left[\sum_{i=j}^{k} (y_i - \hat{y}_m)^2 - \sum_{i=j}^{k} (y_i - \tilde{\hat{y}}_{m+1})^2\right]$$

→ given that $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ is minimizer of $RSS(z) = \sum_{i=1}^{n} (z_i - z)^2$, the new mean of the new buckets $\tilde{\hat{y}}_m$ and $\tilde{\hat{y}}_{m+1}$ minimize the equation $\sum_{i=1} (y_i - \tilde{\hat{y}}_m)^2$ and $\sum_{i=j}^{k} (y_i - \tilde{y}_{m+1})^2$. Therefore;

$$\sum_{i=1}^{j} (y_i - \hat{y}_m)^2 \geq \sum_{i=1}^{j} (y_i - \tilde{\hat{y}}_m)^2 \quad \text{and} \quad \sum_{i=j}^{k} (y_i - \hat{y}_m)^2 \geq \sum_{i=j}^{k} (y_i - \tilde{\hat{y}}_{m+1})^2$$

Given this, it follows that $\sum_{i=1}^{j} (y_i - \hat{y}_m)^2 - \sum_{i=1}^{j} (y_i - \tilde{\hat{y}}_m)^2 \geq 0$

and $\sum_{i=j}^{k} (y_i - \hat{y}_m)^2 - \sum_{i=j}^{k} (y_i - \tilde{\hat{y}}_{m+1}) > 0$

$\therefore \ \Delta = [>0] + [>0] =$

$\therefore \ \Delta \geq 0$

$\Delta =$

1. c) Let R^2_{old} be training set R^2 for T_{old} Model

R^2_{new} = training set R^2 for new model.

$SST = \sum_{i=1}^{N} (y_i - \bar{y})^2$ = total sum square error

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ = overall mean

↳ given value cp $\alpha \geq 0$, modified cost function

$$C_\alpha(T) = C_{imp}(T) + \alpha \cdot SST \cdot |T|$$

↳ show $C_\alpha(T_{New}) \leq C_\alpha(T_{old})$ only if $R^2_{New} - R^2_{old} \geq \alpha$

- $C_\alpha(T_{New}) \leq C_\alpha(T_{old}) \Rightarrow C_\alpha(T_{old}) - C_\alpha(T_{New}) \geq 0$

$$C_\alpha(T_{old}) - C_\alpha(T_{New}) = C_{imp}(T_{old}) - C_{imp}(T_{New}) + \alpha \cdot SST \cdot |T_{old}| - \alpha \cdot SST \cdot |T_{New}|$$

$$C_\alpha(T_{old}) - C_\alpha(T_{New}) = \Delta + \alpha SST (|T_{old}| - |T_{New}|)$$

Note: $R^2 = 1 - \frac{SSE}{SST}$ $\quad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$|T_{old}| = M \quad |T_{New}| = M+1 \quad \therefore |T_{old}| - |T_{New}| = -1$

$\therefore C_\alpha(T_{New}) - C_\alpha(T_{old}) = \Delta + \alpha SST = (SSE_{old} - SSE_{New}) + \alpha SST$

explore if $C_\alpha(T_{New}) \leq C_\alpha(T_{old}) = 0$

$$0 \leq (SSE_{New} - SSE_{old}) - \alpha SST$$

$\therefore \quad \alpha SST \leq SSE_{New} - SSE_{old}$

$$\alpha \leq \frac{SSE_{New} - SSE_{old}}{SST} = \underbrace{\left(1 + \frac{SSE_{New}}{SST}\right)}_{R^2_{New}} - \underbrace{\left(1 + \frac{SSE_{old}}{SST}\right)}_{R^2_{old}}$$

$$\boxed{\alpha \leq R^2_{New} - R^2_{old}}$$