CS484 Homework #5

Exercise 2:

Algorithm: sort y, temp_u = temp_v = temp_us = temp_vs = 0

$Y \leftarrow \sum_{i \in h} y_i$

$Y_s \leftarrow \sum_{i \in h} y_i^2$

for $i = 1$ to n

sort $X_{:,j}$

for $j = 1$ to n

temp_u $+= y_i$

temp_us $+= y_i^2$

temp_v = Y - temp_u

temp_vs = Y_s - temp_us

u_mean = $\frac{\text{temp\_u}}{i}$

v_mean = $\frac{\text{temp\_v}}{n-i}$

loss(i) = temp_us $- 2 \cdot$ u_mean $\cdot$ temp_u $+ i \cdot$ (u_mean)^2 +

temp_vs $- 2 \cdot$ v_mean $\cdot$ temp_v $+ (n-i) \cdot$ (v_mean)^2

end loop

~~min_loss~~

thresh(j) = $\arg\min_i$ loss(i)

end loop

best_thresh = $\min_j$ thresh(j)

best_feature = $\arg\min_j$ thresh(j)

return (best_thresh, best_feature)

correctness:

To solve $\min_j \min_{t_j} \left[\min_{u \in \mathbb{R}} \sum_{i: x_{ij} \le t_j} (y_i - u)^2 + \min_{v \in \mathbb{R}} \sum_{i: x_{ij} > t_j} (y_i - v)^2 \right]$:

if we split elements by a threshold for a specific feature, the value of other features don't matter, so we can consider each feature independtly if we only need threshold for one feature.

Now for any feature selected, we only need to try using each elements as threshold which has total of n.

if we specified a threshold, for any $\min_{u \in \mathbb{R}} \sum_{i: x_{ij} \le t_j} (y_i - u)^2$, the best value would be ~~mean(y_i)~~ mean($\cup y_i$), it holds true for v as well since this will be variance of $\cup y_i$ and the value will increase if u is not the mean.

$$\therefore \min_{u \in \mathbb{R}} \sum_{x_{ij} \le t_j} (y_i - u)^2 + \min_{v \in \mathbb{R}} \sum_{i: x_{ij} > t_j} (y_i - v)^2 = \min_{u \in \mathbb{R}} \sum_{x_{ij} \le t_j} (y_i - \underset{\text{for } u}{\text{mean}}(y_i))^2 + \min_{v \in \mathbb{R}} \sum_{i: x_{ij} > t_j} (y_i - \underset{\text{for } v}{\text{mean}}(\cup y_i))^2$$

$$= \min_{u \in \mathbb{R}} \sum_{x_{ij} \le t_j} y_i^2 - 2 y_i \,\text{mean}(\cup y_i) + (\text{mean}(y_i))^2 +$$

$$\min_{v \in \mathbb{R}} \sum_{x_{ij} > t_j} y_i^2 - 2 y_i \,\text{mean}(\cup y_i) + (\text{mean}(y_i))^2$$

$$= \min_{u \in \mathbb{R}} \sum y_i^2 - 2 \cdot \text{mean}(\cup y_i) \cdot \sum y_i + |\cup y_i| \cdot (\text{mean}(y_i)^2) +$$

$$\min_{v \in \mathbb{R}} \sum y_i^2 - 2\,\text{mean}(\cup y_i) \cdot \sum y_i + (\cup y_i) \cdot \text{mean}(y_i)^2$$

∴ The algorithm solves the equation.

run time: total d features, try n elements for each feature, for each try: sort takes $O(n \log n)$, calculations are ~~linear~~ constant-time.

Final selection: n for each threshold, totally d of them, then choose 1 feature from d

∴ Runtime $T(n) = O(d \cdot (n \log n + a(1))) + O(nd) + O(n) = O(d \cdot n \log n)$