11/17/24 CS483 - Big Data Mining - HW4 Gokulnath Thirumaran ①

67-5086474

1) a) The impurity is

$$I(D) = 100 \times (1-(0.4)^2-(0.6)^2)$$

$$= 100 \times (1-0.16-0.36)$$

$$= 100 \times 0.48$$

$$= 48$$

1.) So, 20 out of 50 wine drinkers like beer

20 out of 50 non-wine drinkers like beer.

Impurity of the "like wine" is $I(D_L)$

$$I(D_L) = 50 \times (1-(0.4)^2-(0.6)^2)$$

$$= 50 \times (1-0.16-0.36)$$

$$= 50 \times 48$$

$$I(D_L) = 24$$

Impurity of the "doesn't like wine" is $I(D_R)$

$$I(D_R) = 50 \times (1-(0.4)^2-(0.6)^2)$$

$$= 50 \times \cancel{2}48$$

$$I(D_R) = 24$$

Thus,

$$G = I(D) - (I(D_L) + I(D_R))$$

$$= 48 - (24 + 24)$$

$$= 48 - 48 = 0.$$

So no reduction in impurity.

2. So,

20 out of 30 runners like beer

20 out of 70 non-runners like beer

The impurity on "like running" side is $I(D_L)$

$$I(D_L) = 30 \times (1 - (0.66)^2 - (0.33)^2)$$

$$= 30 \times (1 - 0.4356 - 0.1089)$$

$$= 30 \times 0.4555$$

$$I(D_L) = 13.665$$

The impurity on "doesn't like running" is $I(D_R)$

$$I(D_R) = 70 \times (1 - (0.2857)^2 - (0.7143)^2)$$

$$= 70 \times (1 - (0.08162) - (0.51022))$$

$$= 70 \times (0.40815)$$

$$I(D_R) = 28.5714$$

Thus,

$$G_1 = I(D) - (I(D_L) + I(D_R))$$
$$= 48 - (13.665 + (28.5714))$$
$$= 48 - 42.2364$$
$$= 5.7636$$

3) So, 50 out of 80 pizza lovers like beer

10 out of 20 non-pizza lovers like beer

The impurity on "like pizza" is $I(D_L)$

$$I(D_L) = 80 \times (1 - (0.375)^2 - (0.625)^2)$$
$$= 80 \times (1 - 0.140625 - 0.390625)$$
$$= 80 \times (1 - 0.53125)$$
$$= 80 \times 0.46875$$
$$I(D_L) = 37.5$$

The impurity on "doesn't like pizza" is $I(D_R)$

$$I(D_R) = 20 \times (1 - (0.5)^2 - (0.5)^2)$$
$$= 20 \times (1 - 0.25 - 0.25)$$
$$= 20 \times (1 - 0.5)$$
$$= 20 \times 0.5 = 10$$
$$I(D_R) = 10$$

Thus,

$$G_1 = I(D) - \left(I(D_L) + I(D_R) \right)$$
$$= 48 - 37.5 - 10$$
$$= 48 - \cancel{48.5}\ 47.5$$
$$= 0.5$$

So, we should choose the running attribute since it has the largest G_1.

b) a_1 will be the root of the tree and left branch denotes $a_1 = 0$, right branch denotes $a_1 = 1$.
The decision tree which avoids overfitting would have a single decision on the root corresponding to a_1; This is because a_1 is a attribute predictive of the outcome, where 1% can be considered as noise, and none of the other attributes are predictive of the outcome.

2) a) Let $T = \{t_z \mid z = 1, \ldots, k\}$. Lets assume that a data point $p \in S_{ij}$ with corresponding clustering center t_{ij} and $p \in S_2$ with corresponding clustering center t_2. By triangular Inequality for Euclidean distance, we have.

$$\|p - t_2\|_2 \leq \|p - t_{ij}\|_2 + \|t_{ij} - t_2\|_2 .$$

Then we square the inequality on both sides,

$$\|p - t_2\|_2^2 \leq \left(\|p - t_{ij}\|_2 + \|t_{ij} - t_2\|_2\right)^2 \leq 2\|p - t_{ij}\|_2^2 + 2\|t_{ij} - t_2\|$$

Then we sum up over all $p \in S_{ij}$, we have

$$\sum_{p \in S_{ij}} \|p - t_2\|_2^2 \leq 2 \sum_{p \in S_{ij}} \|p - t_{ij}\|_2^2 + 2 \sum_{p \in S_{ij}} \|t_{ij} - t_2\|_2^2$$

$$= 2 \sum_{p \in S_{ij}} \|p - t_{ij}\|_2^2 + 2|S_{ij}| \times \|t_{ij} - t_2\|_2^2$$

Suming up j from 1 to k and i from 1 to l,

$$\sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{p \in S_{ij}} \|p - t_2\|_2^2 \leq 2 \sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{p \in S_{ij}} \|p - t_{ij}\|_2^2 +$$

$$2 \sum_{i=1}^{l} \sum_{j=1}^{k} w(t_{ij}) \|t_{ij} - t_2\|_2^2$$

$$\Rightarrow w(t_{ij}) = |S_{ij}|$$

The left side of inequility can be reformulated as,

$$\sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{p \in S_{ij}} \|p - t_2\|_2^2 = \sum_{p \in S} \|p - t_2\|_2^2 = \sum_{p \in S} d(p, T).$$

The right side of the inequality can be reformulated as,

$$2 \sum_{i=1}^{l} \sum_{j=1}^{k} w(t_{ij}) \| t_{ij} - t_2 \|_2^2 + 2 \sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{p \in S_{ij}} \|p - t_{ij}\|_2^2$$

$$= 2 \sum_{t_{ij} \in \hat{S}} w(t_{ij})\, d(t_{ij}, T) + 2 \sum_{i=1}^{l} \sum_{p \in S_i} d(p, T_i)$$

By cost function,

$$\sum_{p \in S} d(p, T) = \mathrm{cost}(S, T)$$

$$\sum_{t_{ij} \in \hat{S}} w(t_{ij})\, d(t_{ij}, T) = \mathrm{cost}_w(\hat{S}, T)$$

$$\sum_{i=1}^{l} \sum_{p \in S_i} d(p, T_i) = \sum_{i=1}^{l} \mathrm{cost}(S_i, T_i)$$

So, we prove that

$$\mathrm{cost}(S, T) \le 2 \cdot \mathrm{cost}_w(\hat{S}, T) + 2 \sum_{i=1}^{l} \mathrm{cost}(S_i, T_i)$$

2) b) Since ALG is an α-approx

$$cost(S_i, \hat{T}_i) \leq \alpha \min_{|T'|=k} \{cost(S_i, T')\} \leq \alpha \, cost(S_i, T^*)$$

Then we sum up i from 1 to l

$$\sum_{i=1}^{l} cost(S_i, \hat{T}_i) \leq \sum_{i=1}^{l} \alpha \cdot cost(S_i, T^*) = \alpha \cdot cost(S, T^*)$$

2) c) Since we use ALG on $\hat{S}$ and ALG is a α-approx algo

$$cost_w(\hat{S}, T) \leq \alpha \cdot \min_{|T'|=k} \{cost(\hat{S}, T')\} \leq \alpha \cdot cost_w(\hat{S}, T^*)$$

Hence, first hint proved.

Let $T^* = \{t_z^* \mid z = 1 \dots k\}$. Suppose a data point $p \in S_{ij}$ with corresponding clustering center t_{ij} and $p \in S_z^*$ with corresponding clustering center t_z^*. By triangular inequality for Euclidean distance,

$$\|t_{ij} - t_z^*\|_2 \leq \|p - t_{ij}\|_2 + \|p - t_z^*\|_2$$

we square the inequalities on both sides,

$$\|t_{ij} - t_2^*\|_2^2 \leq (\|P - t_{ij}\|_2 + \|P - t_2^*\|_2)^2 \leq 2 \cdot \|P - t_{ij}\|_2^2 + 2\|P - t_2^*\|_2^2$$

Then we sum up all $P \in S_{ij}$, j from 1 to k and i from 1 to l,

$$\sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{P \in S_{ij}} \|t_{ij} - t_2^*\|_2^2 \leq 2 \cdot \sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{P \in S_{ij}} \|P - t_{ij}\|_2^2 + 2 \sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{P \in S_{ij}} \|P - t_2^*\|_2^2 .$$

The left side of the inequality can be reformulated,

$$\sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{P \in S_{ij}} \|t_{ij} - t_2^*\|_2^2 = \sum_{i=1}^{l} \sum_{j=1}^{k} |S_{ij}| \cdot \|t_{ij} - t_2^*\|_2^2$$

$$= \sum_{t_{ij} \in \hat{S}} w(t_{ij}) \|t_{ij} - t_2^*\|_2^2$$

$$= \sum_{t_{ij} \in \hat{S}} w(t_{ij}) d(t_{ij}, T^*)$$

The right side of inequality can be reformulated

$$2 \sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{P \in S_{ij}} \|P - t_{ij}\|_2^2 + 2 \cdot \sum_{i=1}^{l} \sum_{j=1}^{k} \sum_{P \in S_{ij}} \|P - t_2^*\|_2^2$$

$$= 2 \sum_{i=1}^{l} \sum_{P \in S_i} d(P, T_i) + 2 \sum_{P \in S} d(P, T^*)$$

By cost function,

$$\sum_{t_{ij} \in S} w(t_{ij})\, d(t_{ij}, T^*) = cost_w(\hat{S}, T^*)$$

$$\sum_{i=1}^{l} \sum_{p \in S_i} d(p, T_i) = \sum_{i=1}^{l} cost(S_i, T_i)$$

$$\sum_{p \in S} d(p, T^*) = cost(S, T^*)$$

So, proving second limit

$$cost_w(\hat{S}, T^*) \le 2 \sum_{i=1}^{l} cost(S_i, T_i) + 2 . cost(S, T^*)$$

By using (a) & (b)

$$cost(S, T) \le 2 . cost_w(\hat{S}, T) + 2 \sum_{i=1}^{l} cost(S_i, T_i)$$

$$\le 2\alpha . cost_w(\hat{S}, T^*) + 2 \sum_{i=1}^{l} cost(S_i, T_i)$$

$$\le 2\alpha \left(2 \sum_{i=1}^{l} cost(S_i, T_i) + 2 . cost(S, T^*)\right) + 2 \sum_{i=1}^{l} cost(S_i; T_i)$$

$$\le 4\alpha . cost(S, T^*) + 4\alpha \sum_{i=1}^{l} cost(S_i, T_i) + 2 \sum_{i=1}^{l} cost(S_i, T_i)$$

$$\le (4d^2 + 6d) \cdot cost(S, T^*) \quad (10)$$

Thus,

$$cost(S, T) \le (4d^2 + 6d) \cdot cost(S, T^*)$$

3) a) $Pr(\tilde{F}[i] \le F[i] + \epsilon t) = 1 - Pr(\tilde{F}[i] \ge F[i] + \epsilon t)$

$$= 1 - Pr\left(\min_j \{ c_{j, h_j(i)} \} \ge F[i] + \epsilon t\right)$$

$$= 1 - Pr\left(c_{j, h_j(i)} \ge F[i] + \epsilon t, \; \forall 1 \le j \le \lceil \log(\tfrac{1}{\delta}) \rceil\right)$$

$$= 1 - \prod_{j=1}^{\lceil \log(\frac{1}{\delta}) \rceil} Pr(c_{j, h_j(i)} \ge F[i] + \epsilon t)$$

The last equality is followed by independence of hash functions.

By Markov's inequality

$$Pr(c_{j, h_j(i)} \ge F[i] + \epsilon t) \le \frac{E[c_{j, h_j(i)} - F[i]]}{\epsilon t}$$

By second property,

$$E\left[C_j h_j(i)\right] \le F[i] + \frac{\epsilon}{e}\left(t - F[i]\right)$$

$$\Leftrightarrow E\left[C_j h_j(i) - F[i]\right] \le \frac{\epsilon}{e}\left(t - F[i]\right) < \frac{\epsilon t}{e}$$

Thus,

$$\Pr\left(C_j h_j(i) \ge F[i] + \epsilon t\right) \le \frac{E\left[C_j h_j(i) - F[i]\right]}{\epsilon t}$$

$$\le \frac{1}{e}$$

Insert $\Pr\left(C_j h_j(i) \ge F[i] + \epsilon t\right) \le \frac{1}{e}$ into 1st equation

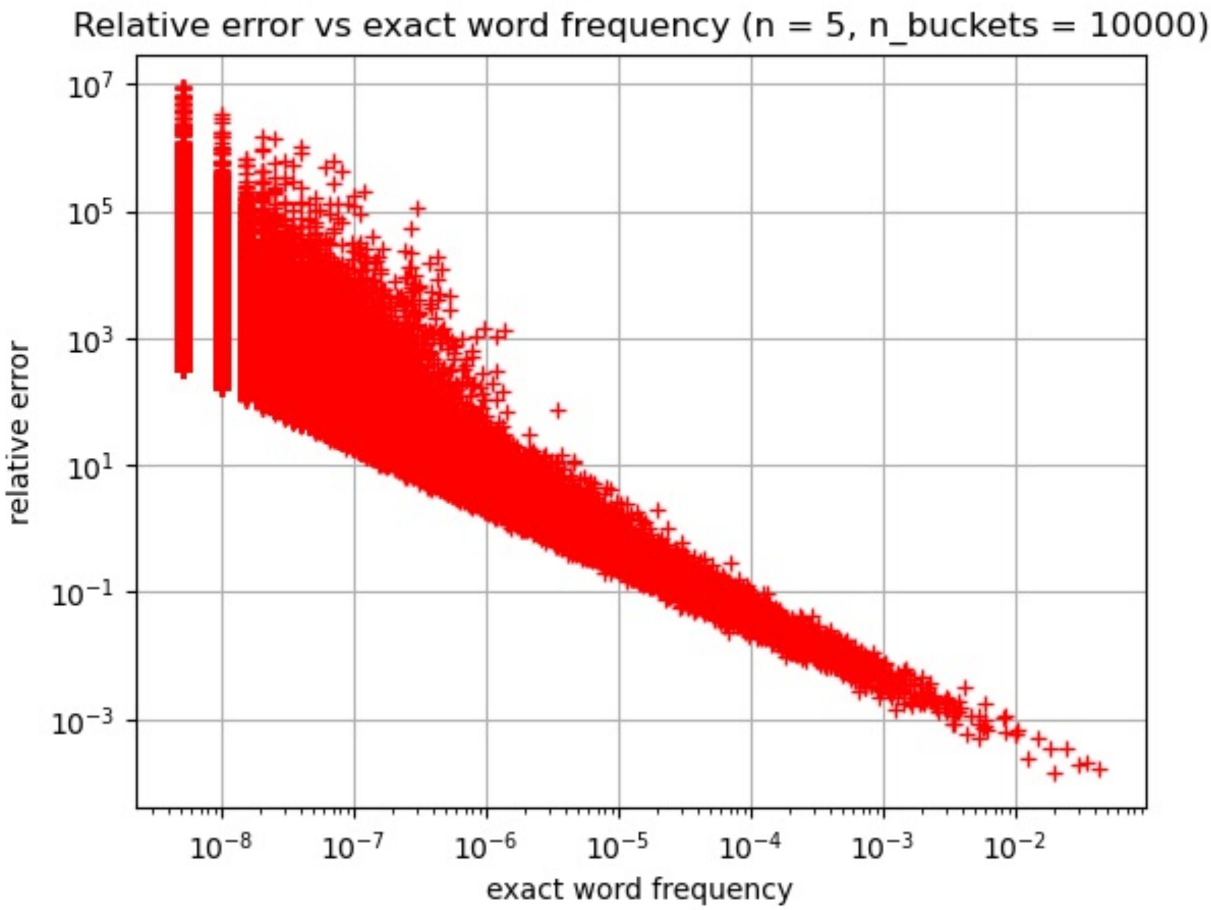
$$\Pr\left(\check{F}[i] \le F[i] + \epsilon t\right) = 1 - \prod_{j=1}^{\lceil \log\left(\frac{1}{\delta}\right) \rceil} \Pr\left(C_j h_j(i) \ge F[i] + \epsilon t\right)$$

$$= 1 - \left(\frac{1}{e}\right)^{\lceil \log\left(\frac{1}{\delta}\right) \rceil} \ge 1 - \delta$$

we prove that

$$\Pr\left(\check{F}[i] \le F[i] + \epsilon t\right) \ge 1 - \delta.$$

3) b)



Relative error vs exact word frequency (n = 5, n_buckets = 10000)
relative error
exact word frequency
10^7
10^5
10^3
10^1
10^{-1}
10^{-3}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}