The cross entropy $L(p,q)$ is defined as $L(p,q) = -\sum_{i=1}^{n} p_i \log q_i$. The gradient of cross entropy loss $L(y,q)$ with respect to $q_i$ is then

$$\begin{aligned}
\frac{\partial L(y,q)}{\partial q_i} &= -\frac{\partial}{\partial q_i} y_i \log(q_i) \\
&= -y_i \frac{1}{q_i} \\
&= -\frac{y_i}{q_i}
\end{aligned}$$

The softmax function is defined as

$$\mu(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

where

$$\mu(x) = \left[\frac{e^{x_1}}{\sum_j e^{x_j}}, \ldots, \frac{e^{x_n}}{\sum_j e^{x_j}}\right]$$

Then the gradient of softmax function $\mu(x)_j$ with respect to $x_i$, observing two different cases, is

- $i = j$

$$\begin{aligned}
\frac{\partial \frac{e^{x_j}}{\sum_k e^{x_k}}}{\partial x_j} &= \frac{e^{x_j} \sum_k e^{x_k} - e^{x_j} e^{x_j}}{(\sum_k e^{x_k})^2} \\
&= \frac{e^{x^j} (\sum_k e^{x_k} - e^{x_j})}{(\sum_k e^{x_k})^2} \\
&= \frac{e^{x^j}}{\sum_k e^{x_k}} \cdot \frac{\sum_k e^{x_k} - e^{x_j}}{\sum_k e^{x_k}} \\
&= \frac{e^{x^j}}{\sum_k e^{x_k}} \cdot \left(\frac{\sum_k e^{x_k}}{\sum_k e^{x_k}} - \frac{e^{x_j}}{\sum_k e^{x_k}}\right) \\
&= \mu(x)_j \cdot (1 - \mu(x)_j)
\end{aligned}$$

- $i \neq j$

$$\begin{aligned}
\frac{\partial \frac{e^{x_j}}{\sum_k e^{x_k}}}{\partial x_i} &= \frac{0 - e^{x_j} e^{x_i}}{(\sum_k e^{x_k})^2} \\
&= -\frac{e^{x_j}}{\sum_k e^{x_k}} \cdot \frac{e^{x_i}}{\sum_k e^{x_k}} \\
&= -\mu(x)_j \cdot \mu(x)_i
\end{aligned}$$

Thus,

$$\frac{\partial \mu(x)_j}{\partial x_i} = \begin{cases} \mu(x)_j \cdot (1 - \mu(x)_j) & \text{if } i = j \\ -\mu(x)_j \cdot \mu(x)_i & \text{if } i \neq j \end{cases}$$

The gradient of fully connected layer $f_{w,b(x)}$ with respect to the $ij$-th entry of weight $W$ and bias $b$ is

$$\begin{aligned}\frac{\partial f_{w,b(x)}}{\partial W_{ij}} &= \frac{\partial (xW+b)}{\partial W_{ij}}\\ &= \frac{\partial (xW)}{\partial W_{ij}} + \frac{\partial b}{\partial W_{ij}}\\ &= x_i\end{aligned}$$

$$\begin{aligned}\frac{\partial f_{w,b(x)}}{\partial b} &= \frac{\partial (xW+b)}{\partial b}\\ &= \frac{\partial (xW)}{\partial b} + \frac{\partial b}{\partial b}\\ &= 1\end{aligned}$$

Not implemented.

In this problem we use Gini index $1 - \sum_i p_i^2$ for impurity measure in binary decision tree. We now define $I(D)$ as

$$I(D) = |D| \times \left(1 - \sum_1 p_i^2\right)$$

and we pick an attribute that maximizes the gain at each node:

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

where $D_L$ and $D_R$ are the sets on left and right hand-site branches after division.
We now calculate the values of $G$ for wine, wunning and pizza attributes (as said in the task).
Before any splitting, the impurity is $I(D) = 100 \times (1 - (0.4^2 + 0.6^2)) = 48$.

- *Wine.*
  - *Impurity on the "likes".* 20 out of 50 who like wine also like beer. It follows

    $$I(D_L) = 50 \times (1 - (0.4^2 + 0.6^2)) = 24$$

  - *Impurity on the "dislikes".* 20 out of 50 who do not like wine like beer. It follows

    $$I(D_D) = 50 \times (1 - (0.4^2 + 0.6^2)) = 24$$

  Thus,

  $$G = 48 - (24 + 24) = 0$$

- *Running.*
  - *Impurity on the "likes".* 20 out of 30 who like running also like beer. It follows

    $$I(D_L) = 30 \times \left(1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right)\right) = 13.333$$

  - *Impurity on the "dislikes".* 20 out of 70 who do not like running like beer. It follows

    $$I(D_D) = 70 \times \left(1 - \left(\left(\frac{2}{7}\right)^2 + \left(\frac{5}{7}\right)^2\right)\right) = 28.5712$$

  Thus,

  $$G = 48 - (13.3333 + 28.5712) = 6.1$$

- *Pizza.*
  - *Impurity on the "likes".* 30 out of 80 who like pizza also like beer. It follows

    $$I(D_L) = 80 \times \left(1 - \left(\left(\frac{3}{8}\right)^2 + \left(\frac{5}{8}\right)^2\right)\right) = 37.5$$

- *Impurity on the "dislikes".* 20 out of 70 who do not like running like beer. It follows

$$I(D_D) = 20 \times \left(1 - \left(0.5^2 + 0.5^2\right)\right) = 10$$

Thus,

$$G = 48 - (37.5 + 10) = 0.5$$

To split the data at the root if we were to maximize the gain $G$ using the Gini index metrix, we would use the running attribute, because it has the highest value of the Gini index. This means that, measuring only how well each single attribute classifies the data set, the running attribute best classifies the data set.

Let us consider the example described in the task.
Let $a_1$ be the root node. We determine the left branch for $a_1 = 0$ and right for $a_1 = 1$. In that case, around 99% of the leaves in the left will be negative, while for the right branch, 99% will be positive. Since we use values of all attributes, each side would consider all of the attributes. Such tree would also avoid overfitting, since the decision at the root is correspoding to $a_1$, on which depend 99% of the target variables of the data points.

First, we show that the cost of the final clustering can be bounded in terms of the total cost of the intermediate clusterings. We prove that $\text{cost}(S,T) \leq 2 \cdot \text{cost}_w(\hat{S},T) + 2\sum_{i=1}^{l} \text{cost}(S_i,T_i)$.

Let $T(x) = \text{arg min}_{t\in T} d(t,x)$. Following the hint we have

$$d(x,T) \leq d(x,t_{ij}) + d(t_{ij},T)$$

for any $x \in S_{ij}$ and $1 \leq i \leq l$, $1 \leq j \leq k$. From tha fact in the task we also know

$$\begin{aligned} d(x,T)^2 &\leq (d(x,t_{ij}) + d(t_{ij},T))^2 \\ &\leq 2d(x,t_{ij})^2 + 2d(t_{ij},T)^2 \end{aligned}$$

We now sum over all elements and get

$$\begin{aligned} \text{cost}(S,T) = \sum_{x\in S_{ij}} \text{cost}(x,T) &= \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x\in S_{ij}} d(x,T)^2 \\ &\leq \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x\in S_{ij}} \left(2d(x,t_{ij})^2 + 2d(t_{ij},T)^2\right) \\ &= \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x\in S_{ij}} 2d(x,t_{ij})^2 + \sum_{i=1}^{l}\sum_{j=1}^{k}\sum_{x\in S_{ij}} 2d(t_{ij},T)^2 \\ &= 2\sum_{i=1}^{l} \text{cost}(S_i,T_i) + \sum_{i=1}^{l}\sum_{j=1}^{k} |S_{ij}| d(t_{ij},T)^2 \\ \Rightarrow \text{cost}(S,T) &\leq 2\sum_{i=1}^{l} \text{cost}(S_i,T_i) + 2\text{cost}_w(\hat{S},T) \end{aligned}$$

We prove that $\sum_{i=1}^{l} \text{cost}(S_i, T_i) \leq \alpha \cdot \text{cost}(S, T^*)$.
Note that $S_1, \ldots, S_l$ form partition of $S$ and $\text{cost}(S_i, T_i) \leq \alpha \cdot \text{cost}(S_i, T_i^*)$, where $T_i^*$ is the optimal clustering for $S_i$ for $1 \leq i \leq l$. Since $T^*$ is optimal clustering for $S$, it follows that

$$\text{cost}(S_i, T_i^*) \leq \text{cost}(S_i, T^*)$$

We thus have

$$\begin{aligned}
\sum_{i=1}^{l} \text{cost}(S_i, T_i) &\leq \alpha \sum_{i=1}^{l} \text{cost}(S_i, T^*) \\
&= \alpha \sum_{i=1}^{l} \sum_{x \in S_i} d(x, T^*)^2 \\
&= \alpha \sum_{x \in S} d(x, T^*)^2 \\
&= \alpha \cdot \text{cost}(S, T^*)
\end{aligned}$$

We now prove that $\text{cost}(S, T) \leq (4\alpha^2 + 6\alpha) \cdot \text{cost}(S, T^*)$.
Following the hint

- we show that $\text{cost}_w(\hat{S}, T) \leq \alpha \text{cost}_w(\hat{S}, T^*)$.

  Let $\hat{T}^*$ be the best clustering for $\hat{S}$, so

  $$\text{cost}_w(\hat{S}, T) \leq \alpha \text{cost}_w(\hat{S}, \hat{T}^*) \quad \text{and} \quad \text{cost}_w(\hat{S}, \hat{T}^*) \leq \text{cost}_w(\hat{S}, T^*)$$

  Thus we have

  $$\text{cost}_w(\hat{S}, T) \leq \alpha \text{cost}_w(\hat{S}, T^*)$$

- we show that $\text{cost}_w(\hat{S}, T^*) \leq 2 \sum_{i=1}^{l} \text{cost}(S_i, T_i) + 2 \cdot \text{cost}(S, T^*)$.

  We use similar arguments as in task 3(a). For each $x \in S_{ij}$ we have

  $$d(t_{ij}, T^*) \leq d(t_{ij}, x) + d(x, T^*)$$

  and

  $$d(t_{ij}, T^*)^2 \leq (d(t_{ij}, x) + d(x, T^*))^2 \leq 2d(t_{ij}, x)^2 + 2d(x, T^*)^2$$

  We now sum over all elements and get the desired inequality

  $$\begin{aligned}
  \text{cost}_w(\hat{S}, T^*) &= \sum_{i=1}^{l} \sum_{j=2}^{k} \sum_{x \in S_{ij}} d(t_{ij}, T^*)^2 = \sum_{i=1}^{l} \sum_{j=2}^{k} |S_{ij}| d(t_{ij}, T^*)^2 \\
  &\leq \sum_{i=1}^{l} \sum_{j=2}^{k} \sum_{x \in S_{ij}} \left(2d(t_{ij}, x)^2 + 2d(x, T^*)^2\right) \\
  &= \sum_{i=1}^{l} \sum_{j=2}^{k} \sum_{x \in S_{ij}} 2d(t_{ij}, x)^2 + \sum_{i=1}^{l} \sum_{j=2}^{k} \sum_{x \in S_{ij}} 2d(x, T^*)^2 \\
  &= 2 \sum_{i=1}^{l} \text{cost}(S_i, T_i) + 2\text{cost}(S, T^*) \\
  \Rightarrow \text{cost}_w(\hat{S}, T^*) &\leq 2 \sum_{i=1}^{l} \text{cost}(S_i, T_i) + 2\text{cost}(S, T^*)
  \end{aligned}$$

Using previous inequalities we have

$$\begin{aligned}
\text{cost}(S,T) &\leq 2\text{cost}_w(\hat{S},T) + 2\sum_{i=1}^{l}\text{cost}(S_i,T_i) \\
&\leq 2\alpha\text{cost}_w(\hat{S},T^*) + 2\alpha\text{cost}(S,T^*) \\
&\leq 2\alpha\left(2\sum_{i=1}^{l}\text{cost}(S_i,T_i) + 2\text{cost}(S,T^*)\right) + 2\alpha\text{cost}(S,T^*) \\
&\leq 2\alpha\left(2\alpha\text{cost}(S,T^*) + 2\text{cost}(S,T^*)\right) + 2\alpha\text{cost}(S,T^*) \\
&= 4\alpha^2\text{cost}(S,T^*) + 4\alpha\text{cost}(S,T^*) + 2\alpha\text{cost}(S,T^*) \\
&= \left(4\alpha^2 + 6\alpha\right)\text{cost}(S,T^*)
\end{aligned}$$

We prove that $\Pr\left[\tilde{F}[i] \le F[i] + \epsilon t\right] \ge 1 - \delta$.
Because $\tilde{F}[i] = \min_j c_{j,h_j(i)}$, $\tilde{F}[i] - F[i] \le \epsilon t$ implies $\min_j c_{j,h_j(i)} - F[i] \le \epsilon t$ for all $1 \le j \le n$. Then we have

$$\Pr\left[\tilde{F}[i] \le F[i] + \epsilon t\right] = 1 - \Pr\left[\tilde{F}[i] \ge F[i] + \epsilon t\right]$$

and

$$\begin{aligned}\Pr\left[\tilde{F}[i] \ge F[i] + \epsilon t\right] &= \Pr\left[\tilde{F}[i] - F[i] \ge \epsilon t\right] \\ &= \prod_{j=1}^{\lceil \log(1/\delta) \rceil} \Pr[c_{j,h_j(i)} - F[i] \ge \epsilon t]\end{aligned}$$

Following the hint (Markov's inequality) we have

$$\Pr[c_{j,h_j(i)} - F[i] \ge \epsilon t] \le \frac{1}{\epsilon t} \cdot \mathbb{E}[c_{j,h_j(i)} - F[i]] \le \frac{1}{e}$$

Thus

$$\begin{aligned}\Pr\left[\tilde{F}[i] - F[i] \ge \epsilon t\right] &= \prod_{j=1}^{\lceil \log(1/\delta) \rceil} \Pr[c_{j,h_j(i)} - F[i] \ge \epsilon t] \\ &\le \left(\frac{1}{e}\right)^{\lceil \log(1/\delta) \rceil} \le \left(\frac{1}{e}\right)^{\log(1/\delta)} = \left(\frac{1}{e}\right)^{-\log \delta} = e^{\log \delta} = \delta\end{aligned}$$

Finally, note that

$$\Pr\left[\tilde{F}[i] \le F[i] + \epsilon t\right] = 1 - \Pr\left[\tilde{F}[i] \ge F[i] + \epsilon t\right] \ge 1 - \delta$$

which finishes the proof.

The code is available in the attached file `HW4_q4.html`.

Log-log plot of the relative error as a function of the frequency is shown on the figure 1.

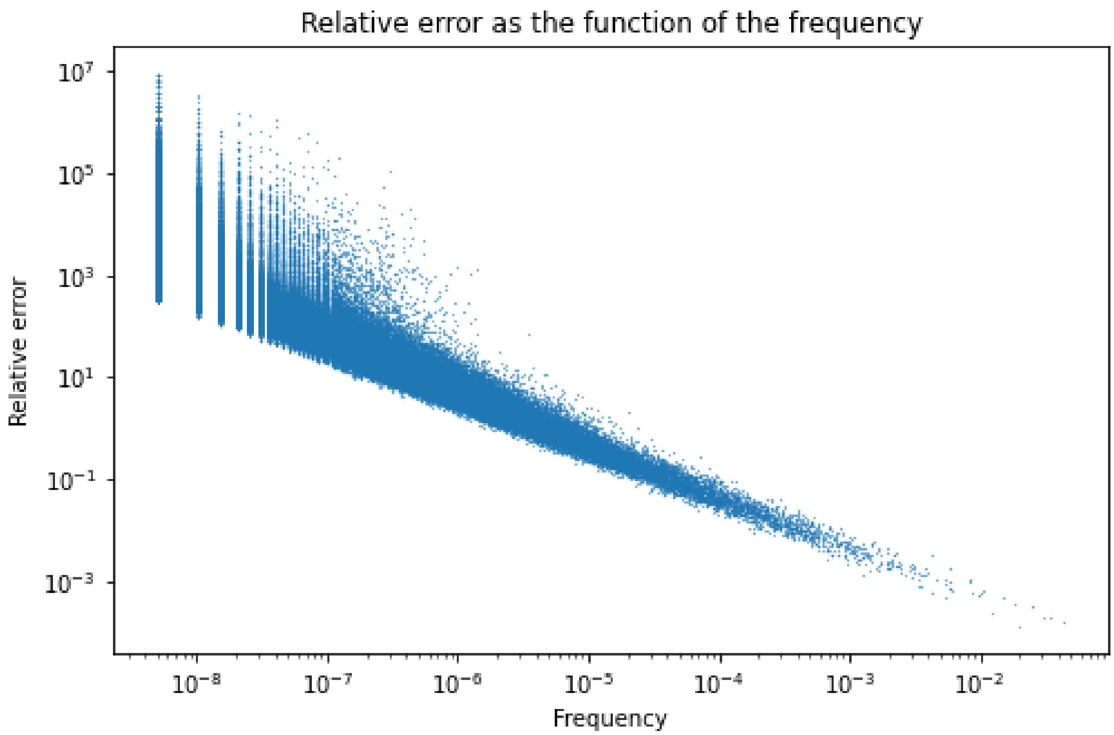


Figure 1: Plot of the relative error as a function of the frequency in log-log scale.

For the relative error to be below 1 (which is $10^0$ on the $y$-axis), the word frequency should be approximately a little bit greater than $10^{-5}$.

# Information sheet
# CS246: Mining Massive Data Sets

**Assignment Submission** Fill in and include this information sheet with each of your assignments. This page should be the last page of your submission. Assignments are due at 11:59pm and are always due on a Thursday. All students (SCPD and non-SCPD) must submit their homework via Gradescope (http://www.gradescope.com). Students can typeset or scan their homework. Make sure that you answer each (sub-)question on a separate page. That is, one answer per page regardless of the answer length. Students also need to upload their code on Gradescope. Put all the code for a single question into a single file and upload it.

**Late Homework Policy** Each student will have a total of *two* late periods. *Homework are due on Thursdays at 11:59pm PT and one late period expires on the following Monday at 11:59pm PT*. Only one late period may be used for an assignment. Any homework received after 11:59pm PT on the Monday following the homework due date will receive no credit. Once these late periods are exhausted, any assignments turned in late will receive no credit.

**Honor Code** We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently, i.e., each student must understand the solution well enough in order to reconstruct it by him/herself. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web (GitHub/Google/previous year's solutions etc.) is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code very seriously and expect students to do the same.

**Your name:** Sara Bizjak ______________________________

**Email:** sarabizjak97@gmail.com __________ **SUID:** 27202020 __________

Discussion Group: Petra Podlogar ______________________________

I acknowledge and accept the Honor Code.

*(Signed)* ______________________________