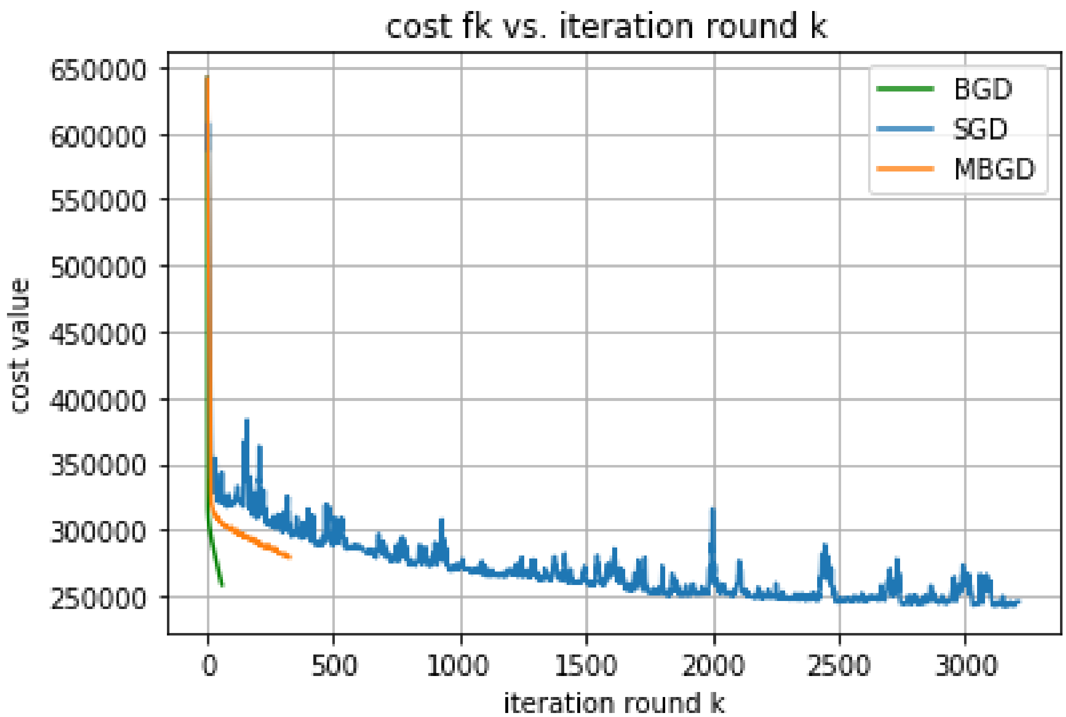
CSE547: Machine Learning for Big Data
Homework 4

Answer to Question 1(a)



Time for each method is 2525s, 675s and 92s, respectively.

Answer to Question 2(a)

I(D) = 100(1-0.75^2 - 0.25^2) = 37.5.

Do split using chocolate ice cream attribute: I(D_L) = 50(1- 0.8^2-0.2^2) = 15 and I(D_R) = 50(1-0.7^2-0.3^2) = 21 and thus G = 1.5
Do split using chocolate ice cream attribute: I(D_L) = 70(1- $(\frac{6}{7})^2$-$(\frac{1}{7})^2$) = 17.1429 and I(D_R) = 30(1-0.5^2-0.5^2) = 15, and thus G = 5.3571
Do split using chocolate ice cream attribute: I(D_L) = 80(1- 0.75^2-0.25^2) and I(D_R)= 20(1-0.75^2-0.25^2) and thus G = 0
Therefore, i would like to use hiking attribute.

Answer to Question 2(b)

a_1 would be at the root of the decision tree (since we assume that the values taken by y depend on a_2,..., a_{100} for fewer than 99 of the samples), and at the left would be the data with $a_1 = 0$ and at the right with $a_1 = 0$, while which attributes at other nodes is not determined.
To avoid overfitting, the left of root (i.e. $a_1 = 0$) should only have one leaf labeled + and the right of root (i.e. $a_1 = 1$) should only have one leaf labeled -, since the 1% data might be noise.

Answer to Question 3(a)

The memory usage is O($\frac{1}{\epsilon} log(\frac{1}{\delta})$).
For each hash function h_j, j$\in$ {1,2,...,[log($\frac{1}{\delta}$)]}, there are [$\frac{e}{\epsilon}$] buckets. And there are [log($\frac{1}{\delta}$)] such hash functions in total.

Answer to Question 3(b)

$\widetilde{F}[i] \geq c_{j,h_j(i)}$ for any j $\in$ {1,2,...,[log($\frac{1}{\delta}$)]}. Since for any i, $h_j(i)$ will hash them to the same buckets, but each bucket of hash function j may have items other than i, thus we have $c_{j,h_j(i)} \geq F[i]$.
Therefore, $\widetilde{F}[i] \geq F[i]$

Answer to Question 3(c)

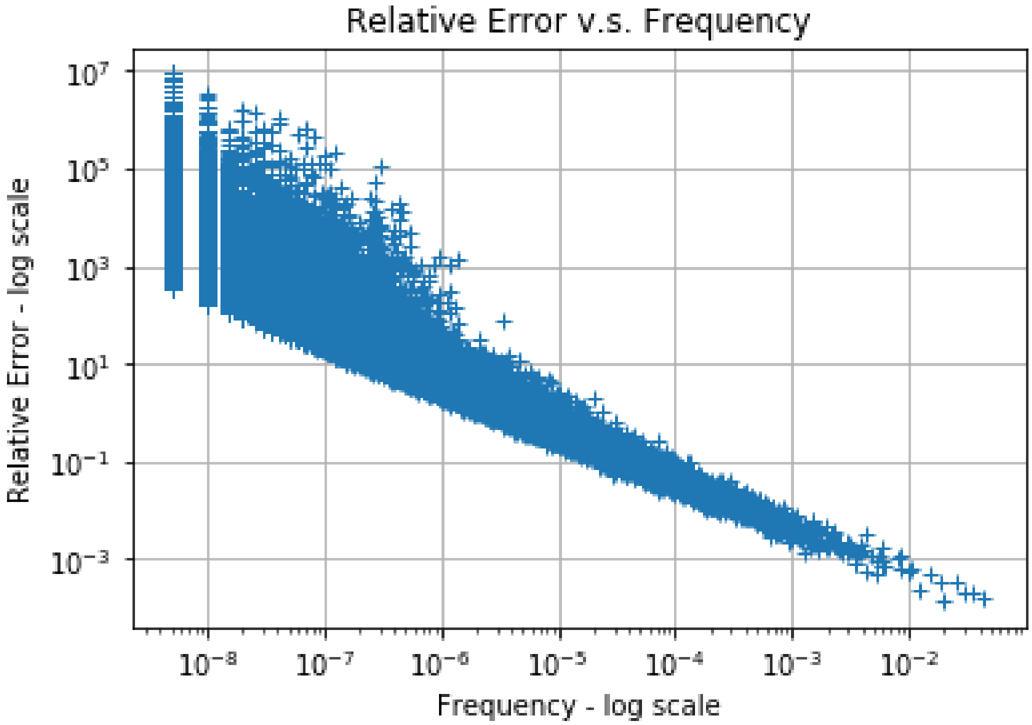
E$[c_{j,h_j(i)}]$
$= F[i]P(c_{j,h_j(i)} = F[i]) + \sum_{c_{j,h_j(i)} > F[i]} c_{j,h_j(i)} P[c_{j,h_j(i)} > F[i]]$
$\leq F[i]\frac{\epsilon}{e}(1-\frac{\epsilon}{e})^{n-1} + \sum_{k=1}^{n-1} t\binom{n-1}{k}(\frac{\epsilon}{e})^{k+1}(1-\frac{\epsilon}{e})^{n-1-k}$
$\leq F[i] + t(\frac{\epsilon}{e})\sum_{k=1}^{n-1}\binom{n-1}{k}(\frac{\epsilon}{e})^{k}(1-\frac{\epsilon}{e})^{n-1-k}$
$\leq F[i] + t(\frac{\epsilon}{e})$
(The third line is because $c_{j,h_j(i)} > F[i]$) when items besides i are hashed to the same bucket, i.e., $h_j(p) = h_j(i)$ for item p. Since there are $\frac{\epsilon}{e}$ values of h_j, thus the prob that k items are hashed to the same buckets as i is $(\frac{\epsilon}{e})\binom{n-1}{k}(\frac{\epsilon}{e})^{k}(1-\frac{\epsilon}{e})^{n-1-k}$)

Answer to Question 3(d)

$Pr[\widetilde{F}[i] \leq F[i] + \epsilon t]$
$= 1 - Pr[\widetilde{F}[i] \geq F[i] + \epsilon t]$
$= 1 - Pr[c_{j,h_j(i)} \geq F[i] + \epsilon t$ for any j]
$= 1 - \prod_j Pr[c_{j,h_j(i)} \geq F[i] + \epsilon t]$

By markov's inequality, $Pr[c_{j,h_j(i)} - F[i] \geq \epsilon t] \leq \frac{E[c_{j,h_j(i)} - F[i]]}{\epsilon t} \leq \frac{1}{e}$
Thus, $Pr[\widetilde{F}[i] \leq F[i] + \epsilon t] \geq 1 - (\frac{1}{e})^{log\frac{1}{\delta}} = 1 - \delta$

Answer to Question 3(e)



From the plot, a word of frequency larger than 10^{-5} tends to have relative error smaller than 1.

Answer to Question 4(a)

Proof: Note that $Z = \sum_j h(j)F[j]$. Then,

$E_h(X) = E_h(Z^2) = E_h[(\sum_j h(j)F[j])^2]$
$= E_h[(\sum_j h(j)^2 F[j])^2 + 2\sum_{i<j} h(i)h(j)F(i)F(j)]$
$= E_h[\sum_j F[j]^2] + E_h[\sum_j h(i)h(j)F(i)F(j)]$
$= \sum_j F[j]^2 + 2F(i)F(j)]E_h[\sum_{j<l} h(i)h(j)]$
$= M + 2F(i)F(j)]E_h[\sum_{j<l} h(i)h(j)]$.

$E_h[\sum_{j<l} h(i)h(j)] = \sum_{j<l} E[h(i)h(j)] = \sum_{j<l}(1^2 * 1/4 + (-1)^2 * 1/4 + (-1) * 1/2) = 0$
Therefore, $E_h(X) = M$

Answer to Question 4(b)

$Var(X) = E(X^2) - E(X)^2$
$= E[(\sum_j h(j)F[j])^4 + \sum_{j_1,j_2,j_3,j_4} h(j_1)h(j_2)h(j_3)h(j_4)F[j_1]F[j_2]F[j_3]F[j_4]] - M^2$
$= E[\sum_j h(j)^4 F[j]^4] + E[\sum_{j_1,j_2,j_3,j_4} h(j_1)h(j_2)h(j_3)h(j_4)F[j_1]F[j_2]F[j_3]F[j_4]] - M^2$

$E[\sum_{j_1,j_2,j_3,j_4} h(j_1)h(j_2)h(j_3)h(j_4)F[j_1]F[j_2]F[j_3]F[j_4]]$
$= F[j_1]F[j_2]F[j_3]F[j_4]E[\sum_{j_1,j_2,j_3,j_4} h(j_1)h(j_2)h(j_3)h(j_4)]$
If j_1, j_2, j_3, j_4 are all different from each other, we have,
$E[h(j_1)h(j_2)h(j_3)h(j_4)] = 2/16 * 1 + 6/16 * 1 + 4/16 * (-1) + 4/16 * (-1) = 0$

If three of j_1, j_2, j_3, j_4 are same and one is different from other three.
$E[h(j_1)h(j_2)h(j_3)h(j_4)] = (-1)*4/16+(-1)*(-1)*4/16+1*(-1)*4/16+(-1)*1*4/16 = 0$

If two of j_1, j_2, j_3, j_4 are same and the left two are different.
$E[h(j_1)h(j_2)h(j_3)h(j_4)] = 1 * 1 * 1/4 + (-1) * (-1) * 1/4 + 1 * (-1) * 1/2 = 0$

Therefore, $E[\sum_{j_1,j_2,j_3,j_4} h(j_1)h(j_2)h(j_3)h(j_4)F[j_1]F[j_2]F[j_3]F[j_4]]$
$= E[\sum_{j_1,j_2,j_3,j_4} h(j_1)h(j_2)h(j_3)h(j_4)F[j_1]F[j_2]F[j_3]F[j_4]]$
$= 6\sum_{i<j} E[h(i)^2h(j)^2F[i]^2F[j]^2]$
$= 6\sum_{i<j} F[i]^2F[j]^2E[h(i)^2h(j)^2]$
$= 6\sum_{i<j} F[i]^2F[j]^2$

Then, Var(X) = $E[\sum_j h(j)^4F[j]^4] + 6\sum_{i<j} F[i]^2F[j]^2 - M^2$
$= \sum_j F[j]^4 + 6\sum_{i<j} F[i]^2F[j]^2 - M^2$
$\leq 3\sum_j F[j]^4 + 6\sum_{i<j} F[i]^2F[j]^2 - M^2 - M^2$
$= 3M^2 - M^2$
$= 2M^2$