CMSC33581 UNIT 1 REVIEW

MAKSIM LEVENTAL

Exercise 1. For some finite set of words S and any collision-resistant hash function h, let *Median Hash* $h_{\text{med}}(S)$ be the median value of the image of S under h. That is to say

$$h_{\text{med}}(S) \coloneqq \text{med}(\{h(s) \mid s \in S\})$$

Problem (1a)**.** Let $\sigma \in \Sigma$ be a uniformly random permutation and S, S' be finite sets of words. Does

$$\mathbb{P}_{\Sigma}[h_{\text{med}}(\sigma(S)) = h_{\text{med}}(\sigma(S'))] = \text{J}(S, S')$$

Solution (1a)**. No**. We prove by contradiction; let

$$S \coloneqq \{\text{'a'}\} \quad S' \coloneqq \{\text{'a'}, \text{'b'}\} \tag{0.1}$$

Clearly $h_{\text{med}}(\sigma(S)) = h(\sigma(\text{'a'}))$ and

$$h_{\text{med}}(\sigma(S')) = \frac{h(\sigma(\text{'a'})) + h(\sigma(\text{'b'}))}{2}$$

for any permutation σ (assuming no collisions). Therefore, since h is collision-resistant

$$\mathbb{P}_{\sigma\in\Sigma}[h_{\text{med}}(\sigma(S)) = h_{\text{med}}(\sigma(S'))] = 0$$

while

$$\text{J}(S, S') = \frac{|\{\text{'a'}\} \cap \{\text{'a'}, \text{'b'}, \text{'b'}\}|}{|\{\text{'a'}\} \cup \{\text{'a'}, \text{'b'}, \text{'b'}\}|} = \frac{|\{\text{'a'}\}|}{|\{\text{'a'}, \text{'b'}\}|} = \frac{1}{2}$$

Problem (1b)**.** Under which conditions would the Median Hash be an exact estimator of Jaccard similarity?

Solution (1b)**.** The crux of the issue is that the median of a set of numbers isn't necessarily in the set:

$$\text{med}(\{1, 2\}) = 1.5$$

Note, further, that

$$\text{med}(\{1, 2\}) = \text{med}(\{0, 3\})$$

and therefore strings with an even number of words might spuriously appear to have high similarity according to median hash.

On the other hand, **if the cardinality of the set is odd** then the median is necessarily a member of the set. Thus, if the set of words (i.e. after making distinct) has odd cardinality then we claim $h_{\text{med}}(S)$ is an estimator for Jaccard similarity. We prove by reduction to Minimum Hash $h_{\min}$ (which is known to accurately estimate Jaccard similarity). First, for any finite set A with odd cardinality,

Date: March 3, 2021.

let h' be the map that maps the minimum of A to the median of A and vice-versa, but otherwise is the identity:

$$h'(a) := \begin{cases} \operatorname{med}(\{a \mid a \in A\}) & \text{if } s = \min(\{a \mid a \in A\}) \\ \min(\{a \mid a \in A\}) & \text{if } s = \operatorname{med}(\{a \mid a \in A\}) \\ a & \text{otherwise} \end{cases}$$

Note that h' is well defined since $\operatorname{med}(\{a\}) \in A$. Then for any hash function h, any finite set of words S with odd cardinality, and any permutation σ

$$(h' \circ h)_{\min}(\sigma(S)) = h_{\text{med}}(\sigma(S))$$

and thus

$$\begin{aligned}\mathbb{P}_{\sigma\in\Sigma}\left[h_{\text{med}}(\sigma(S)) = h_{\text{med}}(\sigma(S'))\right] &= \mathbb{P}_{\sigma\in\Sigma}\left[(h' \circ h)_{\min}(\sigma(S)) = (h' \circ h)_{\min}(\sigma(S'))\right] \\ &= \mathrm{J}(S, S')\end{aligned}$$

since $h' \circ h$ is a valid hash function.

Exercise 2. Let $A := [\text{NOV}-11, \text{DEC}-12, \ldots, \text{NOV}-13, \text{MAR}-02]$. THIS IS ABOUT CARDINALITIES (THINK ABOUT INDEP HASH TABLES)

Problem (2a)**.** Applying standard dictionary encoding to the collection of dates will?

Solution (2a)**.** By the uniqueness of the 100 strings, a dictionary encoding **will never decrease the cardinality** of the dataset because otherwise some strings would be lost.

Problem (2b)**.** Suppose, we calculate an optimal prefix-free code (like Huffman coding) to... the collection of dates will?

Solution (2b)**.** Code words (and therefore code word lengths) will be **arbitrarily associated with each date**. But note they won't have arbitrary lengths because otherwise the code wouldn't be an optimal; we know from Huffman coding that the average code length is approximately $\log_2(100) \approx 6.644$ and therefore at least one code word must have length < 7 and at least one code word must have length ≥ 7 (i.e. not same code word lengths).

Problem (2c)**.** Consider the following encoding algorithm...

Solution (2c)**.** The algorithm produces two dictionaries $D_{\text{MM}}, D_{\text{DD}}$ with

$$O(|\{\text{'MAR'}, \text{'NOV'}, \text{'DEC'}\}|) \leq |D_{\text{MM}}| \leq O(|\{\text{'JAN'}, \ldots, \text{'DEC'}\}|)$$
$$O(|\{\text{'02'}, \text{'11'}, \text{'12'}, \text{'13'}\}|) \leq |D_{\text{DD}}| \leq O(|\{\text{'01'}, \ldots, \text{'31'}\}|)$$

Therefore

$$|D_{\text{MM}}| + |D_{\text{DD}}| \leq 12 + 31 < 100$$

and hence **will always decrease the size of the dataset**.

Exercise 3. Let $S := \{s_1, \ldots, s_N\}$ be the set a N strings. Suppose, you are given a single hash function $\texttt{hash}(s)$ that takes in a string s as an argument and returns an integer from 1 to H and satisfies the simple uniform hashing assumption (SUHA).

Problem (3a)**.** Is it guaranteed that at least two different strings from the list have the same hash code?

Solution (3a)**. True**, by pigeonhole principle.

Problem (3b)**.** Define

```
def h2(s): return hash(str(hash(s)))
```

The maximum that `h2` can return is?

Solution (3b)**. Possibly less than** H. Let $S' \coloneqq \{\text{'1'}, \dots, \text{'H'}\}$ be the string representations for integers $L \coloneqq \{1, \dots, H\}$, with $H < N$, and $j \coloneqq \max\left(\{\texttt{hash}\,(s) \mid s \in S\}\right)$. Note that $j \in L$. If there does not exist $s' \in S'$ such that $\texttt{hash}\,(s') = j$ then `h2` could not possibly attain the same maximum as `hash` over S. Therefore, suppose there exists $s' \in S'$ such that $\texttt{hash}\,(s') = j$ and let l be the integer corresponding to s'. Note that $l \in L$. Then by SUHA, for $s \in S$ we have

$$\mathbb{P}\left[\texttt{hash}\,(s) = l\right] = \frac{1}{H} \Rightarrow \mathbb{P}\left[\texttt{hash}\,(s) \neq l\right] = 1 - \frac{1}{H}$$

and furthermore

$$\begin{aligned}
\mathbb{P}\left[\texttt{hash}\,(s_1) \neq l \wedge \cdots \wedge \texttt{hash}\,(s_N) \neq l\right] &= \prod_{i=1}^{N} \mathbb{P}\left[\texttt{hash}\,(s_i) \neq l\right] \\
&= \prod_{i=1}^{N} \left(1 - \mathbb{P}\left[\texttt{hash}\,(s_i) \neq l\right]\right) \\
&= \prod_{i=1}^{N} \left(1 - \frac{1}{H}\right) \\
&= \left(1 - \frac{1}{H}\right)^N
\end{aligned}$$

and hence

$$\begin{aligned}
\mathbb{P}\left[\max\left(\{\texttt{hash}\,(s) \mid s \in S\}\right) \neq \max\left(\{\texttt{h2}\,(s) \mid s \in S\}\right)\right] &= \mathbb{P}\left[\texttt{hash}\,(s_1) \neq l \wedge \cdots \wedge \texttt{hash}\,(s_N) \neq l\right] \\
&= \left(1 - \frac{1}{H}\right)^N > 0
\end{aligned}$$

i.e. with non-zero probability $\max\left(\{\texttt{h2}\,(s) \mid s \in S\}\right)$ is **possibly less than** $\max\left(\{\texttt{hash}\,(s) \mid s \in S\}\right)$**.**

Problem (3c)**.** Define

```
def h2(s): return hash(s[1:])
```

What applies under the assumptions above?

Solution (3c)**.**

(1) **True**: `h2` is independent from `hash`. Proof: let $z \coloneqq x \cdot z'$ the concatenation of a symbol x and a string z and let $s$be a string such that $s \neq z'$. Then

by SUHA

$$\begin{aligned}\mathbb{P}\left[\mathtt{h2}\left(z\right)=i\wedge\mathtt{hash}\left(s\right)=j\right]&=\mathbb{P}\left[\mathtt{hash}\left(z'\right)=i\wedge\mathtt{hash}\left(s\right)=j\right]\\&=\mathbb{P}\left[\mathtt{hash}\left(z'\right)=i\right]\mathbb{P}\left[\mathtt{hash}\left(s\right)=j\right]\\&=\frac{1}{H}\frac{1}{H}\\&=\mathbb{P}\left[\mathtt{h2}\left(z\right)=i\right]\mathbb{P}\left[\mathtt{hash}\left(s\right)=j\right]\end{aligned}$$

(2) **False**: every possible code does not **necessarily** appear once if you apply `h2` to each of the N strings.
(3) `h2` and `hash` **cannot be combined to make a code space of** H^2 because as in the previous part it's not the case that all code are mapped to by `h2`.
(4) `h2` **is independent from** `hash` **but will no longer satisfy SUHA**: let $z' := x \cdot z$ and $z'' := y \cdot z$ with $x \neq y$. Then

$$\begin{aligned}\mathbb{P}\left[\mathtt{h2}\left(z'\right)=i\wedge\mathtt{h2}\left(z''\right)=j\right]&=\mathbb{P}\left[\mathtt{hash}\left(z\right)=i\wedge\mathtt{hash}\left(z\right)=j\right]\\&=\mathbb{P}\left[\mathtt{hash}\left(z\right)=i\wedge\mathtt{hash}\left(z\right)=j\wedge i\neq j\right]\\&\quad+\mathbb{P}\left[\mathtt{hash}\left(z\right)=i\wedge\mathtt{hash}\left(z\right)=j\wedge i=j\right]\\&=0+\mathbb{P}\left[\mathtt{hash}\left(z\right)=i\right]\\&=\frac{1}{H}\end{aligned}$$

which doesn't equal $1/H^2$, as it should to satisfy SUHA.

Exercise 4. Let P, Q be discrete distributions over $\mathcal{X}$ with $P(x) > 0, Q(x) > 0$ for all $x \in \mathcal{X}$. Prove

$$H(P) \leq \sum_{x\in\mathcal{X}} \frac{P(x)}{Q(x)}$$

Solution (4)**.** Let $H(P)$ be the Shannon entropy of P. Then

$$\begin{aligned}H(P)&=-\sum_{x\in\mathcal{X}}P(x)\log(P(x))\\&\leq-\sum_{x\in\mathcal{X}}P(x)\log(Q(x))\quad\text{by Gibb's inequality}\\&=\sum_{x\in\mathcal{X}}P(x)\log\left(\frac{1}{Q(x)}\right)\\&\leq\sum_{x\in\mathcal{X}}P(x)\left(\frac{1}{Q(x)}-1\right)\quad\text{by }\log(x)\leq x-1\\&=\sum_{x\in\mathcal{X}}\frac{P(x)}{Q(x)}-\sum_{x\in\mathcal{X}}P(x)\\&=\sum_{x\in\mathcal{X}}\frac{P(x)}{Q(x)}-1\leq\sum_{x\in\mathcal{X}}\frac{P(x)}{Q(x)}\end{aligned}$$

Exercise 5. Derive the optimal number of buckets M for a histogram that buckets a discrete random variable X.

Solution. Let $\{0, \ldots, D-1\}$ be the support of X. Then the bins partition the support in $K \coloneqq D/M$-width bins

$$\begin{aligned} &B_1 \coloneqq [0, 1K], B_2 \coloneqq (1K, 2K], \ldots, \\ &\qquad\qquad B_\ell \coloneqq ((\ell-1)K, \ell K], \ldots, B_M \coloneqq ((M-1)K, MK] \end{aligned}$$

and, for a point $x \in B_\ell$, the point mass[1] estimator $\hat{p}_n$ (for $X_1, \ldots, X_n$ i.i.d samples) is defined

$$\hat{p}_n(x) \coloneqq \frac{1}{K}\frac{|\{X_i \in B_\ell\}|}{n} = \frac{1}{K}\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{B_\ell}(X_i)$$

where $\mathbb{1}_A$ is the indicator[2] over event A. Let

$$x^* \coloneqq (\ell-1)K + 1$$

Then the expectiation of $\hat{p}_n(x)$ for $x \in B_\ell$ is

$$\begin{aligned} \mathbb{E}[\hat{p}_n] &= \frac{1}{K}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{1}_{B_\ell}(X_i)] \\ &= \frac{1}{K}\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}[X_i \in B_\ell] \quad \text{by } \mathbb{E}[\mathbb{1}_A] = \mathbb{P}[A] \\ &= \frac{1}{K}\mathbb{P}[X_1 \in B_\ell] \quad \text{by } X_i \text{ i.i.d} \\ &= \frac{1}{K}\sum_{j=(\ell-1)K+1}^{\ell K} p(j) \\ &\leq \frac{1}{K}\left(\sum_{j=0}^{K-1} p(x^*) + j\Delta\right) \quad \text{by } p(j+1) - p(j) \leq \Delta \\ &= \frac{1}{K}\left(Kp(x^*) + \frac{(K-1)K}{2}\Delta\right) \\ &= p(x^*) + \frac{(K-1)}{2}\Delta \end{aligned}$$

Therefore, the bias of $\hat{p}_n(x)$ for $x \in B_\ell$ is

$$\begin{aligned} \text{Bias}[\hat{p}_n] &= \mathbb{E}[\hat{p}_n] - p(x) \\ &= p(x^*) + \frac{(K-1)}{2}\Delta - p(x) \\ &\leq \frac{(K-1)}{2}\Delta + (K-1)\Delta \quad \text{by } |((\ell-1)K+1) - x| \leq K-1 \\ &= \frac{3}{2}(K-1)\Delta \end{aligned}$$

[1] As opposed to density.

[2] If $a \in A$ then $\mathbb{1}_A(a) = 1$ else 0.

Then the variance of $\hat{p}_n(x)$ for $x \in B_\ell$ is

$$\begin{aligned}
\mathrm{Var}[\hat{p}_n] &= \frac{1}{K^2}\,\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{B_\ell}(X_i)\right] \\
&= \frac{1}{K^2}\frac{\mathbb{P}[X_i \in B_\ell](1-\mathbb{P}[X_i \in B_\ell])}{n} \\
&\le \frac{1}{nK^2}\mathbb{P}[X_i \in B_\ell] \quad \text{since } |1-\mathbb{P}[X_i \in B_\ell]| \le 1 \\
&\le \frac{1}{nK^2}\left(Kp(x^*) + \frac{(K-1)K}{2}\Delta\right) \\
&= \frac{p(x^*)}{nK} + \frac{(K-1)}{2nK}\Delta
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{MSE}[\hat{p}_n] &= (\mathrm{Bias}[\hat{p}_n])^2 + \mathrm{Var}[\hat{p}_n] \\
&\le \left(\frac{3}{2}(K-1)\Delta\right)^2 + \frac{p(x^*)}{nK} + \frac{(K-1)}{2nK}\Delta \\
&\le (2K\Delta)^2 + \frac{p(x^*)}{nK} + \frac{\Delta}{n}
\end{aligned}$$

Then solving for the optimal M^{opt}

$$\frac{\partial}{\partial K}\left((2K\Delta)^2 + \frac{p(x^*)}{nK} + \frac{\Delta}{n}\right) = 0 \Rightarrow p(x^*) = 8\Delta^2 K^3 n$$

and therefore

$$K^{\mathrm{opt}} = \frac{1}{2}\left(\frac{p(x^*)}{\Delta^2 n}\right)^{1/3} \Rightarrow M^{\mathrm{opt}} = 2\left(\frac{\Delta^2 D^3 n}{p(x^*)}\right)^{1/3}$$