Lecture 5 – Extensions

Recall:

Any LSH function satisfies the triangle inequality. Proof via

$$\delta_{x,x''} \le \delta_{x x'} + \delta_{x' x''}$$

Lemma: Jacquard mean is a kernel (and so are other simhashes)

Proof: $\delta_{x x'} = \langle e_x, e_{x'} \rangle$ ☐

Lemma: Any $\mathrm{sim}(x, x') = \mathbb{E}_h [\delta_{h(x), h(x')}]$ can be embedded as binary mapping

Proof: Construct new hash function

$$b(h(x)) \in \{0, 1\}$$

$$\mathbb{E}_{h,b} \left[\delta_{b(h(x)), b(h(x'))} \right] = \frac{1}{2} + \frac{1}{2} \mathrm{sim}(x, x')$$

Lemma: not all kernels can be embedded as a sim hash:

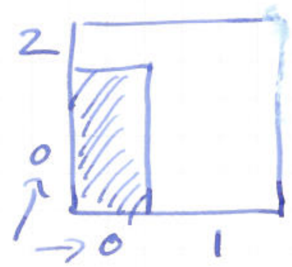
Proof: $\langle (1,0), (-1,0) \rangle = -1$ ↯

Extension $\overline{\mathrm{sim}}(x, x') = \mathbb{E}_{h,b} \{ b(x) b(x') \delta_{h(x), h(x')} \}$

Lemma: $\overline{\mathrm{sim}}(x, x')$ embeds all kernels!

need to ~~app.~~ map $a \cdot b$ to binary

→ sign is easy → $\mathrm{sgn}(a) \to b(i)$



this is the only chance to match

← need absolute upper bound per coordinate

Condition 1) trace class $\sum_i \lambda_i < \infty$

2) L_∞ on eigenbasis $\|\psi_i\|_\infty < \infty$

or rather $\boxed{\sum_i \lambda_i \|\psi_i\|_\infty^2 < \infty}$

but all kernels satisfy this since we have

$$k(x, x') = \sum_i \lambda_i \psi_i(x) \psi_i(x')$$

So, no summand must be ∞ ■

~~Properties of the min-wise hash~~

We can use this to approximate kernels, too via

$$k(x, x') \approx \frac{1}{n} \sum_i \left(h_i(x) = h_i(x') \right) \cdot b_i(x) \, b_i(x')$$

⇒ we need only 2 bit per dimension ✓

Lecture 6 - Properties of the Min-Wise Hash

Min-wise hash (recall)

$$\mathrm{sim}(X, X') = \mathop{\mathbf{E}}_{\pi}\left[\min(\pi(X)) \equiv \min(\pi(X'))\right]$$
$$= \frac{|X \cap X'|}{|X \cup X'|}$$

in practice take $\min_k$ rather than k-times min

Definition:

$$\boxed{\begin{array}{l}\text{min-wise indep. permutation family} \\ \Pr(\min(\pi(X)) = \pi(x)) = \frac{1}{X} \\ \text{for } x \in X\end{array}}$$

Surprising theorem: (Broder & Mitzenmacher, 2001)

For any mapping $f : \mathcal{P}(\mathcal{X}) \to \Omega$ there exists a permutation π_f such that

$$f(X) = f(\pi_f^{-1}(\min \pi_f(X)))$$

and (obviously) there is a distribution over π such that $\mathop{\mathbf{E}}_{\pi}[\min \pi(X) = \min \pi(X')]$

$$= \mathop{\mathbf{E}}_{f}[f(X) = f(X')]$$

Advance Preview:

- f is invertible (to show)
- extract x_i via "f^{-1}"$(X \setminus \{x_1 \dots x_{i-1}\})$

Auxiliary Results

Theorem to prove:

Assume that there is a mapping f, $P(f)$ s.t.

$$\Pr_f\{f(A) = f(B)\} = \frac{|A \cap B|}{|A \cup B|}$$

then there exist some π_f such that

$$f(X) = f(\{\pi_f^{-1}(\min(\pi_f(X)))\})$$

Lemma 1) $\Pr_f(f(X) = f(\{x\})) = \frac{1}{|X|}$ (by definition)

Lemma 2) $\Pr_f(f(\{x\}) = f(\{x'\})) = \delta_{xx'}$ (by definition)

Lemma 3) $\underbrace{\Pr_f(f(X) \in \{f(\{x_i\}) \mid \forall i\})}_{\Downarrow} = 1$

$$\Leftrightarrow \sum_i \Pr_f(f(X) = f(\{x_i\})) = 1$$

(by Lemma 1)

Lemma 4) For $X \subseteq Y$ we have

$$\Pr_f(f(X) = f(Y)) = \frac{|X|}{|Y|}$$

And if $f(Y) \in \{f(x_i) \mid i\}$ then $f(Y) = f(X)$

Proof: Part 1 is trivial

Tail bounds

Corollary: averaging reduces variance

x_i with $\mathbb{E}[x_i] = \mu$

$Var[x_i] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 = \sigma^2$

define $X := \frac{1}{n}\sum_{i=1}^{n} x_i$

$\Rightarrow \mathbb{E}[X] = \mu$; $Var[X] = \frac{1}{n}\sigma^2$

(since variances add up indep.)

Corollary: $Pr\{|X-\mu| > \gamma\sigma\} \leq \frac{1}{\gamma\sqrt{n}}$

→ Terrible scaling behavior in δ, but good in γ, n

→ Want logarithmic in δ, but cannot use Bernstein / Chernoff since the moments are not bounded ... similar to $\min\{x_i\}$ trick

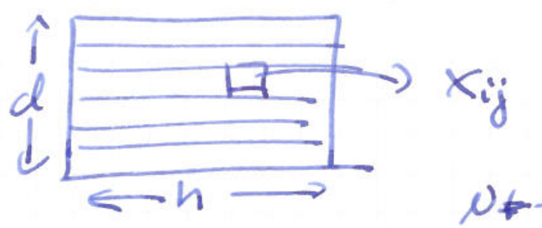
---

Chernoff bound: for $x_i \in [a_i, b_i]$ we have

$$Pr\left\{\sum_i x_i > \sum_i \mathbb{E}[x_i] + \varepsilon\right\} \leq \exp\left(-2\frac{\varepsilon^2}{c^2}\right)$$

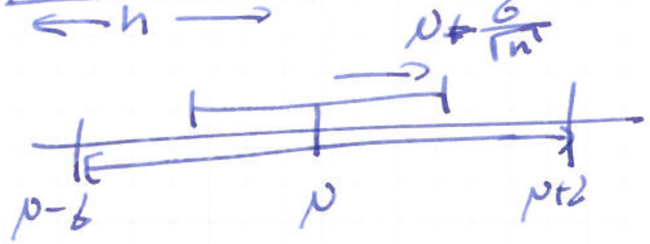
when $c^2 = \sum_i (b_i - a_i)^2$

Key idea: Controlling variance & sample size separately



$\hat{x}_i := \frac{1}{n}\sum_{j=1}^{n} x_{ij}$

$\hat{x} := \text{median}_i\, \hat{x}_i$



$Pr\{|\hat{x}_i - \mu| > \gamma\cdot\frac{\sigma}{\cancel{\sqrt{n}}}\} < \frac{1}{\gamma\sqrt{n}}$

$Pr\{|\text{med}_i\{\hat{x}_i\} - \mu| > \gamma\cdot\sigma\}$

$\leq Pr\left\{\sum_{i=1}^{d}\{|\hat{x}_i - \mu| > \gamma\cdot\sigma\} > \frac{d}{2}\right\}$

Lots of things must go bad before this happens

this is a R.V. with $\mathbb{E}[\ ] \leq \frac{1}{\gamma\sqrt{n}}$

$\Rightarrow \varepsilon = d\left(\frac{1}{2} - \frac{1}{\gamma\sqrt{n}}\right)$

$c^2 = d$

$\Rightarrow Pr\{\text{failure}\} \leq \exp\left(-2d\left(\frac{1}{2} - \frac{1}{\gamma\sqrt{n}}\right)^2\right)$

$\Rightarrow$ set $\frac{1}{\gamma\sqrt{n}} = \frac{1}{4} \Rightarrow \gamma = \frac{1}{4\sqrt{n}}$

$\Rightarrow$ prob is $\exp(-\frac{1}{2}d)$ ✓

McDiarmid ineq. (~95)

$|f(X\setminus\{x_i\} \cup \{x_i'\}) - f(x)| \leq c_i$

$\Rightarrow Pr\{\mathbb{E}[f(x)] - f(x) > \varepsilon\} \leq \exp\left(-2\,\varepsilon^2/c^2\right)$

when $c^2 = \sum_i c_i^2$

Self-bounding ineq of McDiarmid & Reed (~06)

$\sum_i (g(x) - g_i(x)) \leq a\cdot g(x) + b$

where $g_i(x) = \inf_{x'} g(X\setminus\{x_i\} \cup \{x'\})$

then $Pr\{g(x) - \mathbb{E}_x[g(x)] \geq \varepsilon\} \leq e^{-\frac{\varepsilon^2}{2(a\mu+b+a\varepsilon)}}$

$\leq$ — " — $a\varepsilon/3$

min hash details

$$\Pr\{f(x) = f(Y)\}$$
$$= \sum_i \underbrace{\Pr\{f(X) = f(x_i) \mid f(Y) = f(x_i)\}}_{=1} \cdot \underbrace{p(f(Y) = f(\{x_i\}))}_{\frac{1}{|Y|}}$$

Proof of main theorem: (lemma 3)

f is invertible on its image. now define

$x_1 = f^{-1} \circ f(X)$ $\qquad i := \pi_f(x_i)$

$x_i := f^{-1} \circ f(X \setminus \{x_1 \dots x_{i-1}\})$

now we want to apply this to arbitrary sets Y

define $i := \arg\min_i \{x_i \mid x_i \in Y\}$

by construction $Y \in X \setminus \{x_1 \dots x_{i-1}\}$

and $x_i = f^{-1} \circ f(X \setminus \{x_1 \dots x_{i-1}\})$

hence by lemma 4 $f(Y) = f(\{x_i\})$ ∎

(i.e. π_f satisfies the conditions)

Useful result: (will not prove this)

$$\Pr\{\operatorname{rank}_r(\pi(X)) = \pi(x)\} = \frac{1}{|X|}$$

for min-wise independent families.

Tail bounds:

Recall: Gauss-Markov inequality:

$$\Pr(X > \mu c) \le \frac{1}{c} \quad \text{for } X \ge 0$$

Useful corollary

$$\Pr\{(\min_{i \in 1..k} x_i) > \mu c\} \le \left(\frac{1}{c}\right)^k$$

(will use this for Count Min sketch)

Quantile trick

define $F(\bar{x}) = \int_{-\infty}^{\bar{x}} dp(x)$

hence for $x \sim p(x)$ we have $F(x) \sim U[0,1]$

$$\boxed{\begin{aligned} \text{define } F_k(\bar{x}) &:= \Pr\{\min_{i \in 1..k} x_i \le \bar{x}\} \\ &= F^k(\bar{x}) \end{aligned}}$$

Corollary 59/95 trick

define $\bar{x}_{0.95} = F^{-1}(0.95)$

then $\Pr\{\max_{i \in 1..59} x_i < \bar{x}_{0.95}\} < 0.05$

Proof: $(0.95)^{59} < 0.05$ ∎

Chebyshev inequality:

Define $z := (x - \mathbb{E}[x])^2$; $\mathbb{E}[z] = \sigma^2$

$$\Rightarrow \underbrace{\Pr\{z > \gamma^2 \sigma^2\}}_{\Pr\{|x - \mathbb{E}[x]| > \gamma\sigma\}} \le \frac{1}{\gamma^2}$$