MS&E 317/CS 263: Algorithms for Modern Data Models, Spring 2014
`http://msande317.stanford.edu`.
Instructors: Ashish Goel and Reza Zadeh, Stanford University.

Lecture 7, 4/23/2014. Scribed by Simon Anastasiadis.

Cauchy Distribution

The Cauchy distribution has PDF given by:

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

for $x \in (-\infty, \infty)$.

Note that $\int \frac{1}{1+x^2} dx = \arctan(x)$.

The Cauchy distribution has infinite mean and variance.

$$E(|x|) = \frac{2}{\pi}\int_0^\infty \frac{x}{1+x^2}dx = \frac{1}{\pi}\int_0^\infty \frac{1}{1+y}dy = \frac{1}{\pi}\log(1+y)|_0^\infty$$

Normal Random Variables

The normal distribution behaves well under addition.

Consider $z_1 \sim N(0; v_1)$ and $z_2 \sim N(0; v_2)$ are two independent random variables, and scalar a. Then:

$$\begin{aligned} z_1 + z_2 &\sim N(0; v_1 + v_2) \\ az_1 &\sim N(0; a^2 v_1) \end{aligned}$$

Or more generally, for fixed but arbitrary real numbers $r_1, ..., r_k$ and random variables $z_1, ..., z_k$ all iid $N(0;1)$ we have:

$$\begin{aligned} r_1 z_1 + ... + r_k z_k &=_D N(0, r_1^2 + ... + r_k^2) \\ &=_D \sqrt{r_1^2 + ... + r_k^2} N(0;1) \end{aligned}$$

So

$$\begin{aligned} \big(\sum r_i z_i\big)^2 &=_D \Big(N(0; \sum r_i^2)\Big)^2 \\ &=_D \big(\sum r_i^2\big)\big(N(0;1)\big)^2 \end{aligned}$$

P-Stability

A distribution f is p-stable if for any $K > 0$, any real numbers $r_1, ..., r_K$, and given K iid random variables $z_1, ..., z_K$ following distribution f, then

$$\sum_{i=1}^K r_i z_i =_D \big(\sum_{i=1}^K |r_i|^p\big)^{1/p} \ Y$$

where Y has distribution f.

Notes:

- For any $p \in (0, 2]$ there exists some p-stable distribution.
- The Cauchy distribution is 1-stable.
- The Normal distribution is 2-stable.
- The CLT suggests that no other distribution is 2-stable

F2 Estimation

$$F_2(t) = \sum_{a \in U} \left(f_t(a)\right)^2$$

This looks similar to computing a variance.

Define the consistent normal random variable $h_i(a) \sim N(0; 1)$ such that $h_i(a)$ and $h_j(b)$ are independent if $i \neq j$ or $a \neq b$.

In practice:

```
h(i,a) = { rseed(<i,a>)
           return rand_normal()
         }
```

Then for a multiset S (a set where the same element may occur multiple times) we define the following sketch:

$$\sigma =< \sum_{a \in S} h_1(a), ..., \sum_{a \in S} h_J(a) >$$

With τ being coordinate wise addition:

$$\tau(\sigma(S_1), \sigma(S_2) =< ..., \sum_{a \in S_1} h_i(a) + \sum_{a \in S_2} h_i(a), ... >$$

Now given $\sigma =< \sigma_1, ..., \sigma_J >$ we wish to estimate F_2. Notice that:

$$\sigma_i = \sum_{a \in S} h_i(a) = \sum_{a \in U} h_i(a) f_s(a)$$

$$\sigma_i =_D \left(\sum_{a \in U} \left(f_s(a)\right)^2 \right)^{1/2} N(0; 1)$$

Then:

$$\sigma_i^2 =_D \sum_{a \in U} \left(f_s(a)\right)^2 \left(N(0; 1)\right)^2$$

So

$$\text{mean}(\sigma^2) = F_2$$
$$\text{median}(\sigma^2) = 0.4705 F_2$$

Where the number 0.4705 arises as the median of a chi-squared distribution with 1 degree of freedom.

F2 Estimation in Stream

Suppose at time t you have a sketch $< \sigma_1, ..., \sigma_J >$. You then wish to update once a_t arrives according to:

$$< \sigma_1 + h_1(a_t), ..., \sigma_J + h_J(a_t) >$$

Suppose now your data is arranged in key-value pairs $< a_t, v_t >$ such that $f_t(a) = \sum_{t:a=a_t} v_t$. Then the update becomes:

$$< \sigma_1 + v_t h_1(a_t), ..., \sigma_J + v_t h_J(a_t) >$$

F1 Estimation

In the simple case $F_1(t) = t$.

However, when the data is arranged in key-value pairs $< a_t, v_t >$ where the v_t may take on negative values, computing $F_1(t)$ is no longer trivial.

Consider the following example:

$$< 1, -3 > \ , \ < 1, 7 > \ , \ < 2, 1 > \ , \ < 3, -1 > \ , \ < 2, 2 >$$

Produces

$$f_t(1) = 4 \quad f_t(2) = 3 \quad f_t(3) = -1$$

This has $F_1(t) = \sum_i |f_t(i)| = 8$.

To estimate $F_1(t)$ we follow the same technique as for $F_2(t)$ but replacing the consistent normal random variable by a consistent Cauchy random variable. So now let $h_i(a) \sim$ Cauchy. Then:

$$\sigma_1 = \sum_{a \in U} h_1(a) f_t(a) =_D F_1(t)\ C$$

So

$$\text{median}(|\sigma_1|, ..., |\sigma_J|) =_D F_1(t)\ \text{median}(|C_1|, ..., |C_J|)$$

Where C and C_i are Cauchy distributed random variables.

$|C_i|$ has PDF $\frac{2}{\pi} \frac{1}{1+x^2}$ for $x \in [0, \infty)$. The median can be calculated by solving:

$$\frac{1}{2} = \frac{2}{\pi} \int_0^x \frac{1}{1+t^2} dt$$

This gives solution $x = \tan(\frac{\pi}{4}) = 1$.

Therefore the median of σ is a good estimator for $F_1(t)$.

Machine Learning Application

Suppose you have N points in $\mathbb{R}^D$ where N and D are large, and you want to produce a summary or reduction of these points while preserving distance: $d(x, y) = \|X - Y\|_2$.

1. Pick D random variables $z_1, ..., z_D$ iid for $N(0; 1)$. Fix these values.

2. Map $X =< x_1, ..., x_D >$ to $\sum_{i=1}^{D} x_i z_i$. The resulting value will follow a normal distribution multiplied by a constant.

Then for two points X and Y we have:

$$\sum_{i=1}^{D} x_i z_i - \sum_{i=1}^{D} y_i z_i = \sum_{i=1}^{D} (x_i - y_i) z_i \sim N(0;1) \|X - Y\|_2$$

So the L2-norm is (approximately) preserved.