Harvard Privacy Tools Project

Variance Sensitivity Proofs

March 23, 2020

Definition 1. *Let sample variance be defined as*

$$s^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

where $\bar{x}$ refers to the sample mean of x.

1 Neighboring Definition: Change One

1.1 ℓ_1-sensitivity

Lemma 1. *For arbitrary a,*

$$\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(a - \bar{x})^2.$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}\left((x_i - \bar{x}) - (a - \bar{x})\right)^2 \\
&= \sum_{i=1}^{n}\left((x_i - \bar{x})^2 - 2(x_i - \bar{x})(a - \bar{x}) + (a - \bar{x})^2\right) \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\sum_{i=1}^{n}\left(x_i a - x_i\bar{x} - \bar{x}a + \bar{x}^2\right) + \sum_{i=1}^{n}\left(a^2 - 2a\bar{x} + \bar{x}^2\right) \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2a\sum_{i=1}^{n}x_i + 2\bar{x}\sum_{i=1}^{n}x_i + 2\bar{x}an - 2\bar{x}^2 n + a^2 n - 2a\bar{x}n + \bar{x}^2 n \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + a^2 n - 2a\bar{x}n + \bar{x}^2 n \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(a - \bar{x})^2
\end{aligned}$$

$\square$

Theorem 1. *Let*

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Then for $\boldsymbol{x}$ bounded between m and M, $f(\cdot)$ has ℓ_1-sensitivity in the change-one model bounded above by

$$\frac{n-1}{n}(M-m)^2.$$

Proof. Consider databases $\mathbf{x}'$ and $\mathbf{x}''$ which differ in a single point. For notational ease, call $\mathbf{x}$ the part of $\mathbf{x}'$ and $\mathbf{x}''$ that is the same, and say that $\mathbf{x}$ contains n points. WLOG say that the last data point in the database is the one that differs. I.e., $\mathbf{x}' = \mathbf{x} \cup \{x_{n+1}\}$, and $\mathbf{x}'' = \mathbf{x} \cup \{x'_{n+1}\}$. This proof assumes that a "neighboring database" is one that differs in a single data-point, so we will ultimately be comparing $f(\mathbf{x}')$ and $f(\mathbf{x}'')$. However, it is useful to first write $f(\mathbf{x}')$ in terms of $f(\mathbf{x})$. Note that

$$\begin{aligned} \bar{x}' &= \frac{1}{n+1}\sum_{i=1}^{n+1} x_i \\ &= \frac{n\bar{x} + x_{n+1}}{n+1}. \end{aligned} \tag{1.1}$$

Then,

$$\begin{aligned} f(\mathbf{x}') &= \sum_{i=1}^{n}(x_i - \bar{x}')^2 + (x_{n+1} - \bar{x}')^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x}' - \bar{x})^2 + (x_{n+1} - \bar{x}')^2 && \text{(By Lemma 1)} \\ &= f(\mathbf{x}) + n\left(\frac{n\bar{x} + x_{n+1}}{n+1} - \bar{x}\right)^2 + \left(x_{n+1} - \frac{n\bar{x} + x_{n+1}}{n+1}\right)^2 && \text{(By Equation 1.1)} \\ &= f(\mathbf{x}) + n\left(\frac{x_{n+1} - \bar{x}}{n+1}\right)^2 + \left(\frac{nx_{n+1} - n\bar{x}}{n+1}\right)^2 \\ &= f(\mathbf{x}) + (x_{n+1} - \bar{x})^2 \frac{n + n^2}{(n+1)^2} \\ &= f(\mathbf{x}) + (x_{n+1} - \bar{x})^2 \frac{n}{n+1} \end{aligned} \tag{1.2}$$

Now, to bound the sensitivity of f, note that

$$\begin{aligned} \left|f(\mathbf{x}') - f(\mathbf{x}'')\right| &= \left|(x_{n+1} - \bar{x})^2 \frac{n}{n+1} - (x'_{n+1} - \bar{x})^2 \frac{n}{n+1}\right| \qquad (1.3) \\ &\leq (M-m)^2 \frac{n}{n+1}. \end{aligned}$$

The bound in the final line follows from the case where $x_{n+1} = M$ (resp. m) and $\bar{x} = x'_{n+1} = m$ (resp. M).

So we have a bound on the sensitivity of $f(\cdot)$ for a data set of size $n+1$. Traditionally we consider sensitivities on a data set of size n. Redefining $n+1$ as n in the above equation gives

$$(M-m)^2\frac{n-1}{n}.$$

$\square$

Corollary 1. *Variance has ℓ_1-sensitivity in the change-one model bounded above by*

$$(M-m)^2\frac{n-1}{n^2}.$$

Proof. This follows from Theorem 1 with a renormalization by n. $\square$

Corollary 2. *Sample variance has ℓ_1-sensitivity in the change-one model bounded above by*

$$\frac{(M-m)^2}{n}.$$

Proof. This follows from Theorem 1 with a renormalization by $n-1$. $\square$

1.2 ℓ_2-sensitivity

Theorem 2. *Let X be a data set with n elements, $x_1,\ldots,x_n$ and let*

$$f(X)=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

be the sample variance. For X bounded between m and M, $f(\cdot)$ has an ℓ_2-sensitivity in the change-one model of

$$\left(\frac{n-1}{n}(M-m)^2\right)^2$$

Proof. We can pick up from statement 1.3, switching from an ℓ_1 to an ℓ_2 norm and interpreting the data sets in question to be of size n rather than $n+1$.

$$\begin{aligned}
(f(x')-f(x''))^2 &= \left((x_n-\bar{x})^2\frac{n-1}{n}-(x'_n-\bar{x})^2\frac{n-1}{n}\right)^2\\
&= \left(\frac{n-1}{n}\right)^2\left((x_n-\bar{x})^2-(x'_n-\bar{x})^2\right)^2\\
&\leq \left(\frac{n-1}{n}\right)^2\left((M-m)^2\right)^2\\
&= \left(\frac{n-1}{n}\right)^2(M-m)^4\\
&= \left(\frac{n-1}{n}(M-m)^2\right)^2.
\end{aligned}$$

$\square$

Corollary 3. *Variance has ℓ_2-sensitivity in the change-one model bounded above by*

$$\left(\frac{n-1}{n^2}(M-m)^2\right)^2.$$

Proof. This follows from Theorem 2 with a renormalization by n. $\square$

Corollary 4. *Sample variance has ℓ_2-sensitivity in the change-one model bounded above by*

$$\left(\frac{1}{n}(M-m)^2\right)^2.$$

Proof. This follows from Theorem 2 with a renormalization by $n-1$. $\square$

2 Neighboring Definition: Add/Drop One

2.1 ℓ_1-sensitivity

Theorem 3. *Let X be a data set with n elements, $x_1, \ldots, x_n$ and*

$$f(X) = \sum_{i=1}^{n}(x_i - \bar{X})^2.$$

For X bounded between m and M, $f(\cdot)$ has a global ℓ_1-sensitivity in the add/drop one model of

$$\frac{n}{n+1}(M-m)^2$$

Proof. We must consider both adding and removing an element from X.

Adding an element:
Let $X' = X \cup x'_{n+1}$. Recall from Eq. 1.2 that for

$$f(x) = \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$f(x') = f(x) + (x_{n+1} - \bar{x})^2 \frac{n}{n+1}.$$

So,

$$\begin{aligned} \left|f(x') - f(x)\right| &= \left|(x_{n+1} - \bar{x})^2 \frac{n}{n+1}\right| \\ &= \left|(x_{n+1} - \bar{x})^2 \frac{n}{n+1}\right| \\ &\leq (M-m)^2 \, \frac{n}{n+1} \end{aligned} \tag{2.1}$$

Removing an element:
Let $X' = X \setminus \{x_n\}$. Then, rewriting Eq. 1.2 with n set to $n+1$ since "x" in this case is the greater set,

$$f(x) = f(x') + (x_n - \bar{x}')^2 \frac{n-1}{n}.$$

Then,

$$\begin{aligned} \left|f(x) - f(x')\right| &= \left|(x_n - \bar{x}')^2 \frac{n-1}{n}\right| \\ &\leq (M-m)^2 \frac{n-1}{n}, \end{aligned} \tag{2.2}$$

$$\tag{2.3}$$

Note that for any $n \geq 1$,

$$\frac{n}{n+1} > \frac{n-1}{n}. \tag{2.4}$$

So, the worst-case bound always occurs in the "add-one" case, and the ℓ_1-sensitivity of $f(\cdot)$ is in general bounded by

$$(M-m)^2 \frac{n}{n+1}. \tag{2.5}$$

□

Corollary 5. *Variance has ℓ_1-sensitivity in the add/drop one model bounded above by*

$$(M-m)^2 \frac{1}{n+1}.$$

Proof. This follows from Theorem 3 with a renormalization by n. □

Corollary 6. *Sample variance has ℓ_1-sensitivity in the add/drop one model bounded above by*

$$(M-m)^2 \frac{n}{n^2-1}.$$

Proof. This follows from Theorem 3 with a renormalization by $n-1$. □

2.2 ℓ_2-sensitivity

Theorem 4. *Variance has ℓ_2-sensitivity in the add/drop one model bounded above by*

$$\left((M-m)^2 \frac{1}{n+1}\right)^2.$$

Proof. Because the bounds in Equations 2.1 and 2.2 are positive, they hold for their square. Then, by the inequality for $n \geq 1$ in Equation 2.4 it follows that the ℓ_2 sensitivity of the variance is bounded by

$$\left((M-m)^2 \frac{1}{n+1}\right)^2,$$

where the change in constant from Equations 2.1 to $1/(n+1)$ comes from the $1/n$ in the definition of variance. □

Corollary 7. *Sample variance has ℓ_2-sensitivity in the add/drop one model bounded above by*

$$\left((M-m)^2 \frac{n}{n^2-1}\right)^2.$$

Proof. The logic here is the same as for 4 with a renormalization by $n-1$ rather than by n. □