Variance Sensitivity Proof

Ira Globus-Harris

October 9, 2019

Definition 1. *Let variance be defined as*

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

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 *has sensitivity 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

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

Then,

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

Now, to bound the sensitivity of f, note that

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

Now, usually we're interested in sensitivities in terms of the total number of values in the database, which here is $n+1$. So, redefining n as $n+1$ in the above equation gives

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

□

Corollary 1. *Sample variance has sensitivity bounded above by*

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