CSCI 145 Problem 9 Solution: The Center Under Four Losses

Solution

The four costs select modes, medians, the mean, and the midpoint of the extreme readings, respectively. We prove that each minimizes its cost below.

1. Counting Mismatches: Modes

For any candidate $c \in \mathbb{R}$, let $N(c) = \#\{i : y_i = c\}$ be the number of readings equal to c. Then

$$\|\mathbf{r}(c)\|_0 = n - N(c).$$

Minimizing the mismatch count is therefore exactly maximizing the number of matches. Every observed value has at least one match, while an unobserved value has none. Thus the complete set of minimizers consists of the observed values with greatest frequency, called the *sample modes*. Several values may tie for greatest frequency. If all readings are distinct, every observed value is a minimizer. Values between readings are not.

2. Absolute Deviations: Medians

Write the sorted readings as $z_1 \le \cdots \le z_n$. This reordering does not change the sum of absolute errors. Pair the smallest reading with the largest, the next-smallest with the next-largest, and so on. For each pair, the triangle inequality gives

$$|z_j - c| + |z_{n+1-j} - c| \ge z_{n+1-j} - z_j,$$

with equality exactly when $c \in [z_j, z_{n+1-j}]$.

If $n = 2q$ is even, the q pair intervals are nested, and their intersection is $[z_q, z_{q+1}]$. Every c in this interval attains the lower bound for every pair simultaneously. Outside the interval, the middle pair has strictly greater cost, while no other pair can have smaller cost. Therefore the entire closed interval is the minimizing set.

If $n = 2q + 1$ is odd, the q pairs leave one middle reading unpaired. The choice $c = z_{q+1}$ attains every pair's lower bound and makes the remaining term $|z_{q+1} - c|$ zero. Any other choice makes that term strictly positive and cannot reduce the paired terms. Thus the middle reading is the unique minimizer. This argument also covers $n = 1$, when there are no pairs.

Together, the complete sets of minimizers are

$$\arg\min_{c \in \mathbb{R}} \|\mathbf{r}(c)\|_1 = \begin{cases} \{z_{q+1}\}, & n = 2q + 1, \\ [z_q, z_{q+1}], & n = 2q. \end{cases}$$

These are exactly the sample medians. Repeated middle readings can make the even-sample interval a single point. Dividing the objective by n gives MAE and preserves this minimizing set.

3. Euclidean Distance: the Mean

Squaring is strictly increasing on nonnegative numbers, so minimizing $\|\mathbf{r}(c)\|_2$ is equivalent to minimizing its square. Dividing that square by $n > 0$ gives MSE, again without changing the minimizers.

Let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ be the sample mean. Insert and subtract $\bar{y}$ in each residual and expand:

$$\begin{aligned}\sum_{i=1}^{n}(y_i - c)^2 &= \sum_{i=1}^{n}\big((y_i - \bar{y}) + (\bar{y} - c)\big)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + 2(\bar{y} - c)\sum_{i=1}^{n}(y_i - \bar{y}) + n(\bar{y} - c)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - c)^2.\end{aligned}$$

The cross term vanishes because $\sum_i (y_i - \bar{y}) = \sum_i y_i - n\bar{y} = 0$. The first term is independent of c, and the second is nonnegative and zero exactly when $c = \bar{y}$. Thus the mean is the unique minimizer.

4. Worst Deviation: the Midrange

Let $a = \min_i y_i$ and $b = \max_i y_i$ be the smallest and largest readings. For any candidate c, the triangle inequality gives

$$b - a \le |b - c| + |c - a| \le 2\max_i |y_i - c| = 2\|\mathbf{r}(c)\|_\infty.$$

Every candidate therefore has worst deviation at least $(b - a)/2$. The midpoint $m = (a + b)/2$, called the *midrange*, attains this bound: every reading lies in $[a, b]$, so every distance to m is at most $(b - a)/2$, with equality at the endpoints.

To attain the same bound, any minimizing c must satisfy both $|b - c| \le (b - a)/2$ and $|a - c| \le (b - a)/2$. These imply

$$c \ge b - \frac{b - a}{2} = m, \qquad c \le a + \frac{b - a}{2} = m.$$

Hence $c = m$ is the unique minimizer, including when all readings coincide.

Comment on Sensitivity

Cost	Minimizing Center	Determined By
$\|\mathbf{r}(c)\|_0$	Sample modes	Exact repetitions
$\|\mathbf{r}(c)\|_1$	Sample medians	Middle order positions
$\|\mathbf{r}(c)\|_2$	Sample mean	Every reading's magnitude
$\|\mathbf{r}(c)\|_\infty$	Midrange	The two extreme readings

Moving one reading by $\Delta \in \mathbb{R}$ moves the mean by Δ/n. With the other readings fixed and $n \ge 2$, moving one reading arbitrarily far also moves the midrange without bound because that reading becomes an extreme.

For $n \ge 3$, once one moving reading is beyond all the fixed readings, the median set is determined entirely by the fixed readings and stops changing. For $n = 2$, the minimizing interval has the moving

reading as an endpoint, so some medians can move without bound. For $n = 1$, every cost is uniquely minimized by the sole reading itself.

A mode can ignore an isolated distant reading when another value remains strictly more frequent. But exact ties matter: if all readings are distinct, every reading, including a distant one, minimizes the mismatch count. Small perturbations can also break repeated values apart and change the minimizing set. Thus the mode does not provide a general guarantee of resistance to a faulty reading.