Tutorial 2

Introduction to Data Analysis Techniques

1 Least Squares Error

Example 1.1 (Comparing Two Straight-Line Fits)**.** [eg:straight-line-squared-error] Consider the three observations

i	0	1	2
$x(i)$	0	1	2
$y(i)$	1	2	2

Let $c_A, c_B : \mathbb{R}^2$ have coordinates

$$(c_A(0), c_A(1)) = (1, 1), \qquad (c_B(0), c_B(1)) = \left(1, \frac{1}{2}\right)$$

1. State n and m, and write the functions generated by c_A and c_B
2. Define $S \equiv \text{PolyError}(n)(m)(x)(y)$, state its signature, and expand $S(c)$ into a sum of squared differences
3. Calculate $S(c_A)$ and $S(c_B)$. Which candidate line has the smaller sum of squared errors?
4. Use the least-squares derivative criterion to find a coefficient vector $c^* : \mathbb{R}^2$ minimizing S over all straight-line models, and write its fitted function

1. $n = 3$ and $m = 1$. The functions generated by c_A and c_B are:

$$f_A(t) = c_A(0) + c_A(1) \cdot t = 1 + 1 \cdot t = t + 1$$

$$f_B(t) = c_B(0) + c_B(1) \cdot t = 1 + \frac{1}{2} \cdot t = \frac{t}{2} + 1$$

2. $S \equiv \text{PolyError}(n)(m)(x)(y)$

$$\begin{aligned} \text{PolyError} : \Pi n : \mathbb{N}_1.\, \Pi m : \mathbb{N}.\, (\text{Fin}(n) \to \mathbb{R}) \\ \to (\text{Fin}(n) \to \mathbb{R}) \to (\mathbb{R}^{m+1} \to \mathbb{R}) \end{aligned}$$

$$\begin{aligned} n : \mathbb{N}_1 = 3, \qquad m : \mathbb{N} = 1 \\ \vdash \text{PolyError}(3)(1) : (\text{Fin}(3) \to \mathbb{R}) \to (\text{Fin}(3) \to \mathbb{R}) \to (\mathbb{R}^2 \to \mathbb{R}) \end{aligned}$$

   Notice that $x, y : \text{Fin}(3) \to \mathbb{R}$, so we know that

$$S = \text{PolyError}(3)(1)(x)(y) : \mathbb{R}^2 \to \mathbb{R}$$

3. For any coefficient vector $c : \mathbb{R}^2$,

$$\begin{aligned} S(c) &= \text{SquaredError}(3)(x)(y)(\text{PolynomialModelGen}(1)(c)) \\ &= \sum_0^2 \left(i \mapsto d(y(i), \text{PolynomialModelGen}(1)(c)(x(i)))^2\right) \\ &= \sum_0^2 \left(i \mapsto |y(i) - \big(c(0) + c(1) \cdot x(i)\big)|^2\right) \end{aligned}$$

For $c_A = (1, 1)$,

$$\begin{aligned} S(c_A) &= \sum_0^2 \left(i \mapsto |y(i) - 1 - x(i)|^2\right) \\ &= |1 - 1 - 0|^2 + |2 - 1 - 1|^2 + |2 - 1 - 2|^2 \\ &= 0^2 + 0^2 + (-1)^2 \\ &= 1 \end{aligned}$$

For $c_B = \left(1, \frac{1}{2}\right)$,

$$\begin{aligned} S(c_B) &= \sum_0^2 \left(i \mapsto |y(i) - 1 - \frac{1}{2} \cdot x(i)|^2\right) \\ &= |1 - 1 - \frac{1}{2} \cdot 0|^2 + |2 - 1 - \frac{1}{2} \cdot 1|^2 + |2 - 1 - \frac{1}{2} \cdot 2|^2 \\ &= 0^2 + \left(\frac{1}{2}\right)^2 + 0^2 \\ &= \frac{1}{4} \end{aligned}$$

Since $S(c_B) = \frac{1}{4} < 1 = S(c_A)$, candidate B has the smaller sum of squared errors

4. Construct

$$S \equiv c \mapsto (c(0) - 1)^2 + (c(0) + c(1) - 2)^2 + (c(0) + 2 \cdot c(1) - 2)^2$$

We require $\mathrm{D}_0(S)(c) = \mathrm{D}_1(S)(c) = 0$. Differentiating gives

$$\begin{aligned} \mathrm{D}_0(S)(c) &= 2 \cdot (c(0) - 1) + 2 \cdot (c(0) + c(1) - 2) \\ &\quad + 2 \cdot (c(0) + 2 \cdot c(1) - 2) \\ &= 2 \cdot (3 \cdot c(0) + 3 \cdot c(1) - 5), \\ \mathrm{D}_1(S)(c) &= 2 \cdot (c(0) + c(1) - 2) + 4 \cdot (c(0) + 2 \cdot c(1) - 2) \\ &= 2 \cdot (3 \cdot c(0) + 5 \cdot c(1) - 6) \end{aligned}$$

Setting both derivatives to zero and dividing by 2 gives

$$\begin{cases} 3 \cdot c(0) + 3 \cdot c(1) = 5 \\ 3 \cdot c(0) + 5 \cdot c(1) = 6 \end{cases}$$

Subtracting the first equation from the second gives $2 \cdot c(1) = 1$, so $c(1) = \frac{1}{2}$. Substituting into the first equation gives $c(0) = \frac{7}{6}$. Thus the minimizing vector and its fitted function are

$$c^* = \left(\frac{7}{6}, \frac{1}{2}\right), \qquad \mathrm{PolynomialModelGen}(1)(c^*) = t \mapsto \frac{7}{6} + \frac{1}{2} \cdot t$$

Its squared error is $S(c^*) = \frac{1}{6} < \frac{1}{4} = S(c_B)$