8.1 - Discrete Least Squares Approximation

Motivation

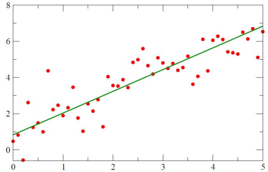

Figure: Dot: measured data; line: linear fit

Q:

Given m data point, $\{(x_i, y_i)\}_{i=1}^{m}$, how find the equation of straight line $y = a_0 + a_1 x$ which is the "best" fit for the data points.

Least squares error:

$$E \equiv E_2(a_0, a_1) = \sum_{i=1}^{m} [y_i - (a_0 + a_1 x_i)]^2$$

Linear least squares method

Ans. to Q: Solve a minimization problem

Find a_0, a_1, such that the least squares error $E_2(a_0, a_1)$ is minimum, i.e., $\min_{a_0,a_1} E_2(a_0, a_1)$

At minimum of error, we have $\frac{\partial E}{\partial a_0} = 0, \text{ and } \frac{\partial E}{\partial a_1} = 0$. This means:

$$0 = \frac{\partial E}{\partial a_0} = \frac{\partial}{\partial a_0} \sum_{i=1}^{m} [y_i - (a_0 + a_1 x_i)]^2 = 2 \sum_{i=1}^{m} [y_i - (a_0 + a_1 x_i)](-1); \quad (1)$$

$$0 = \frac{\partial E}{\partial a_1} = \frac{\partial}{\partial a_1} \sum_{i=1}^{m} [y_i - (a_0 + a_1 x_i)]^2 = 2 \sum_{i=1}^{m} [y_i - (a_0 + a_1 x_i)](-x_i). \quad (2)$$

Simplifying Eq. (1) $\Rightarrow a_0 m + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$;
Simplifying Eq. (2) $\Rightarrow a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i$

These lead to:

Normal equations

$$a_0 m + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \ ;$$

$$a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i \ .$$

$$a_0 = \frac{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} x_i y_i \sum_{i=1}^{m} x_i}{m \sum_{i=1}^{m} x_i^2 - \left(\sum_{i=1}^{m} x_i\right)^2} \tag{3}$$

$$a_1 = \frac{m \sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i}{m \sum_{i=1}^{m} x_i^2 - \left(\sum_{i=1}^{m} x_i\right)^2} \tag{4}$$

Example 1

Find the least squares line approximating the data in the table.

i	x_i	y_i
1	1	1.3
2	2	3.5
3	3	4.2
4	4	5.0
5	5	7.0
6	6	8.8
7	7	10.1
8	8	12.5
9	9	13.0
10	10	15.6

Solution:

i	x_i	y_i	x_i^2	x_iy_i
1	1	1.3	1	1.3
2	2	3.5	4	7.0
3	3	4.2	9	12.6
4	4	5.0	16	20.0
5	5	7.0	25	35.0
6	6	8.8	36	52.8
7	7	10.1	49	70.7
8	8	12.5	64	100.0
9	9	13.0	81	117.0
10	10	15.6	100	156.0

$\sum_{i=1}^{10} x_i = 55$; $\sum_{i=1}^{10} y_i = 81$; $\sum_{i=1}^{10} x_i^2 = 385$; $\sum_{i=1}^{10} x_iy_i = 572.4$;

Polynomial Least squares

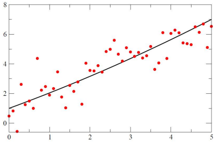

Figure: Dot: measured data; curve: $1 + x + x^2/25$

Q:

Given m data point, $\{(x_i, y_i)\}_{i=1}^m$, find the equation of polynomial: $P_n(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0$ which "best" fits for the data points. Here $n < m - 1$.

Ans to Q:

The coefficients $\mathbf{a} = (a_0, ..., a_n)^t$ are found by minimizing the least squares error $E \equiv E_2(a_0, ..., a_n) = \sum_{i=1}^m [y_i - P_n(x_i)]^2$

$$
\begin{aligned}
E &= \sum_{i=1}^{m} y_i^2 - 2\sum_{i=1}^{m} P_n(x_i)y_i + \sum_{i=1}^{m}(P_n(x_i))^2 \\
&= \sum_{i=1}^{m} y_i^2 - 2\sum_{j=0}^{n} a_j \left(\sum_{i=1}^{m} y_i x_i^j\right) + \sum_{j=0}^{n}\sum_{k=0}^{n} a_j a_k \left(\sum_{i=1}^{m} x_i^{j+k}\right)
\end{aligned} \qquad (5)
$$

At minimum of error, we have $\frac{\partial E}{\partial a_0} = 0, \frac{\partial E}{\partial a_1} = 0, ..., \frac{\partial E}{\partial a_n} = 0$. These lead to:

Normal equations for polynomial fitting

$$
\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k} = \sum_{i=1}^{m} y_i x_i^j, \quad \text{for each } j = 0, 1, ..., n. \qquad (6)
$$

Normal equations in matrix form

$$A^t A\mathbf{a} = A^t\mathbf{b} \tag{7}$$

where

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 & \ldots & x_m^n \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}; \tag{8}$$

and $\mathbf{a} = (a_0, ..., a_n)^t$.

Remark: Eq. (7) has a unique solution if all x_i are unique.

Example. Fit the data in the table using quadratic polynomial least squares method.

i	x_i	y_i
1	0	1.0000
2	0.25	1.2480
3	0.50	1.6487
4	0.75	2.1170
5	1.00	2.7183

Soln: Let the quadratic polynomial be $P_2(x) = a_2x^2 + a_1x + a_0$. The matrix A and vector $\mathbf{b}$ of the normal equation (7) are:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0.25 & 0.25^2 \\ 1 & 0.5 & 0.5^2 \\ 1 & 0.75 & 0.75^2 \\ 1 & 1.00 & 1.0^2 \end{bmatrix}; \quad \mathbf{b} = \begin{bmatrix} 1.0 \\ 1.2840 \\ 1.6487 \\ 2.1170 \\ 2.7183 \end{bmatrix} \tag{9}$$

$$A^t = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0.25 & 0.5 & 0.75 & 1.0 \\ 0 & 0.0625 & 0.25 & 0.5625 & 1 \end{bmatrix}$$

$$A^tA = \begin{bmatrix} 5.0 & 2.5 & 1.875 \\ 2.5 & 1.875 & 1.5625 \\ 1.875 & 1.5625 & 1.3828 \end{bmatrix}; \quad A^t\mathbf{b} = \begin{bmatrix} 8.768 \\ 5.4514 \\ 4.4015375 \end{bmatrix}$$

The normal equations are:

$$\begin{aligned} 5.0a_0 + 2.5a_1 + 1.875a_2 &= 8.768 \\ 2.5a_0 + 1.875a_1 + 1.5625a_2 &= 5.4514 \\ 1.875a_0 + 1.5625a_1 + 1.3828a_2 &= 4.4015375 \end{aligned} \tag{10}$$

Solving Eq. (10) for a_0, a_1, a_2 yields $P_2(x) = 1.0051 + 0.8642x + 0.8437x^2$.
Remark: See matlab code for solving this problem as well.