Least squares problems

The least squares problem (simplest version)

Definition 1

Given a collection of points $(x_1, y_1), \ldots, (x_n, y_n)$, find the coefficients a, b of the line $y = a + bx$ such that

$$\|\mathbf{e}\| = \sqrt{\varepsilon_1^2 + \cdots + \varepsilon_n^2} = \sqrt{(y_1 - \tilde{y}_1)^2 + \cdots + (y_n - \tilde{y}_n)^2}$$

is minimal, where $\tilde{y}_i = a + bx_i$ for $i = 1, \ldots, n$

We just saw how to solve this by brute force using a genetic algorith to minimise $\|e\|$, let us now see how to solve this problem "properly"

For a data point $i = 1, \ldots, n$

$$\varepsilon_i = y_i - \tilde{y}_i = y_i - (a + bx_i)$$

So if we write this for all data points,

$$\begin{aligned} \varepsilon_1 &= y_1 - (a + bx_1) \\ &\vdots \\ \varepsilon_n &= y_n - (a + bx_n) \end{aligned}$$

In matrix form

$$\mathbf{e} = \mathbf{b} - A\mathbf{x}$$

with

$$\mathbf{e} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, A = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \mathbf{x} = \begin{pmatrix} a \\ b \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

The least squares problem (reformulated)

Definition 2 (Least squares solutions)

Consider a collection of points $(x_1, y_1), \ldots, (x_n, y_n)$, a matrix $A \in \mathcal{M}_{mn}$, $\mathbf{b} \in \mathbb{R}^m$. A **least squares solution** of $A\mathbf{x} = \mathbf{b}$ is a vector $\tilde{\mathbf{x}} \in \mathbb{R}^n$ s.t.

$$\forall \mathbf{x} \in \mathbb{R}^n, \quad \|\mathbf{b} - A\tilde{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|$$

Needed to solve the problem

Definition 3 (Best approximation)

Let V be a vector space, $W \subset V$ and $\mathbf{v} \in V$. The **best approximation** to $\mathbf{v}$ in W is $\tilde{\mathbf{v}} \in W$ s.t.

$$\forall \mathbf{w} \in W, \mathbf{w} \neq \tilde{\mathbf{v}}, \quad \|\mathbf{v} - \tilde{\mathbf{v}}\| < \|\mathbf{v} - \mathbf{w}\|$$

Theorem 4 (Best approximation theorem)

Let V be a vector space with an inner product, $W \subset V$ and $\mathbf{v} \in V$. Then $\mathrm{proj}_W(\mathbf{v})$ is the best approximation to $\mathbf{v}$ in W

Let us find the least squares solution

$\forall \mathbf{x} \mathbb{R}^n$, $A\mathbf{x}$ is a vector in the **column space** of A (the space spanned by the vectors making up the columns of A)

Since $\mathbf{x} \in \mathbb{R}^n$, $A\mathbf{x} \in \text{col}(A)$

$\implies$ least squares solution of $A\mathbf{x} = \mathbf{b}$ is a vector $\tilde{\mathbf{y}} \in \text{col}(A)$ s.t.

$$\forall \mathbf{y} \in \text{col}(A), \quad \|\mathbf{b} - \tilde{\mathbf{y}}\| \leq \|\mathbf{b} - \mathbf{y}\|$$

This looks very much like Best approximation and Best approximation theorem

Putting things together

We just stated: The least squares solution of $A\mathbf{x} = \mathbf{b}$ is a vector $\tilde{\mathbf{y}} \in \mathrm{col}(A)$ s.t.

$$\forall \mathbf{y} \in \mathrm{col}(A), \quad \|\mathbf{b} - \tilde{\mathbf{y}}\| \leq \|\mathbf{b} - \mathbf{y}\|$$

We know (reformulating a tad):

Theorem 5 (Best approximation theorem)

Let V be a vector space with an inner product, $W \subset V$ and $\mathbf{v} \in V$. Then $\mathrm{proj}_W(\mathbf{v}) \in W$ is the best approximation to $\mathbf{v}$ in W, i.e.,

$$\forall \mathbf{w} \in W, \mathbf{w} \neq \mathrm{proj}_W(\mathbf{v}), \quad \|\mathbf{v} - \mathrm{proj}_W(\mathbf{v})\| < \|\mathbf{v} - \mathbf{w}\|$$

$\implies W = \mathrm{col}(A)$, $\mathbf{v} = \mathbf{b}$ and $\tilde{\mathbf{y}} = \mathrm{proj}_{\mathrm{col}(A)}(\mathbf{b})$

So if $\tilde{\mathbf{x}}$ is a least squares solution of $A\mathbf{x} = \mathbf{b}$, then

$$\tilde{\mathbf{y}} = A\tilde{\mathbf{x}} = \text{proj}_{\text{col}(A)}(\mathbf{b})$$

We have

$$\mathbf{b} - A\tilde{\mathbf{x}} = \mathbf{b} - \text{proj}_{\text{col}(A)}(\mathbf{b}) = \text{perp}_{\text{col}(A)}(\mathbf{b})$$

and it is easy to show that

$$\text{perp}_{\text{col}(A)}(\mathbf{b}) \perp \text{col}(A)$$

So for all columns $\mathbf{a}_i$ of A

$$\mathbf{a}_i \cdot (\mathbf{b} - A\tilde{\mathbf{x}}) = 0$$

which we can also write as $\mathbf{a}_i^T(\mathbf{b} - A\tilde{\mathbf{x}}) = 0$

For all columns $\mathbf{a}_i$ of A,

$$\mathbf{a}_i^T(\mathbf{b} - A\tilde{\mathbf{x}}) = 0$$

This is equivalent to saying that

$$A^T(\mathbf{b} - A\tilde{\mathbf{x}}) = \mathbf{0}$$

We have

$$\begin{aligned} A^T(\mathbf{b} - A\tilde{\mathbf{x}}) = \mathbf{0} &\iff A^T\mathbf{b} - A^TA\tilde{\mathbf{x}} = \mathbf{0} \\ &\iff A^T\mathbf{b} = A^TA\tilde{\mathbf{x}} \\ &\iff A^TA\tilde{\mathbf{x}} = A^T\mathbf{b} \end{aligned}$$

The latter system constitutes the **normal equations** for $\tilde{\mathbf{x}}$

Least squares theorem

Theorem 6 (Least squares theorem)

$A \in \mathcal{M}_{mn}$, $\mathbf{b} \in \mathbb{R}^m$. *Then*

1. $A\mathbf{x} = \mathbf{b}$ *always has at least one least squares solution* $\tilde{\mathbf{x}}$
2. $\tilde{\mathbf{x}}$ *least squares solution to* $A\mathbf{x} = \mathbf{b} \iff \tilde{\mathbf{x}}$ *is a solution to the normal equations* $A^T A\tilde{\mathbf{x}} = A^T\mathbf{b}$
3. A *has linearly independent columns* $\iff A^T A$ *invertible.*
   In this case, the least squares solution is unique and

$$\tilde{\mathbf{x}} = \left(A^T A\right)^{-1} A^T\mathbf{b}$$

We have seen 1 and 2, we will not show 3 (it is not hard)

Suppose we want to fit something a bit more complicated..

For instance, instead of the affine function

$$y = a + bx$$

suppose we want to do the quadratic

$$y = a_0 + a_1 x + a_2 x^2$$

or even

$$y = k_0 e^{k_1 x}$$

How do we proceed?

Fitting the quadratic

We have the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and want to fit

$$y = a_0 + a_1 x + a_2 x^2$$

At (x_1, y_1),

$$\tilde{y}_1 = a_0 + a_1 x_1 + a_2 x_1^2$$

$\vdots$

At (x_n, y_n),

$$\tilde{y}_n = a_0 + a_1 x_n + a_2 x_n^2$$

In terms of the error

$$\varepsilon_1 = y_1 - \tilde{y}_1 = y_1 - (a_0 + a_1 x_1 + a_2 x_1^2)$$
$$\vdots$$
$$\varepsilon_n = y_n - \tilde{y}_n = y_n - (a_0 + a_1 x_n + a_2 x_n^2)$$

i.e.,

$$\mathbf{e} = \mathbf{b} - A\mathbf{x}$$

where

$$\mathbf{e} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Theorem 6 applies, with here $A \in \mathcal{M}_{n3}$ and $\mathbf{b} \in \mathbb{R}^n$

Fitting the exponential

Things are a bit more complicated here

If we proceed as before, we get the system

$$y_1 = k_0 e^{k_1 x_1}$$
$$\vdots$$
$$y_n = k_0 e^{k_1 x_n}$$

$e^{k_1 x_i}$ is a nonlinear term, it cannot be put in a matrix

However: take the ln of both sides of the equation

$$\ln(y_i) = \ln(k_0 e^{k_1 x_i}) = \ln(k_0) + \ln(e^{k_1 x_i}) = \ln(k_0) + k_1 x_i$$

If $y_i, k_0 > 0$, then their ln are defined and we're in business..

$$\ln(y_i) = \ln(k_0) + k_1 x_i$$

So the system is

$$\mathbf{y} = A\mathbf{x} + \mathbf{b}$$

with

$$A = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \mathbf{x} = \begin{pmatrix} k_1 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} \ln(k_0) \end{pmatrix} \text{ and } \mathbf{y} = \begin{pmatrix} \ln(y_1) \\ \vdots \\ \ln(y_n) \end{pmatrix}$$