Neural Computation

28 September 2023

Linear regression: toy example

- Commute time on bus
  - Want to predict commute time to University
  - Input variables (features)?
    - Distance to University
    - Day of the week
  - Output / target?
    - Commute time
  - Data

Dist (km)	Commute time (min)
2.7	25
4.1	33
1.0	15
5.2	45
2.8	22

$$f(\boldsymbol{x}) = x \ w$$

Linear regression: least squares (1d)

- Let $d = 1$, then:

$$C(w) = \frac{1}{2n} \sum_{i=1}^{n} \left(y^i - x^i w\right)^2$$

How important is this factor?

	x	y
	Dist (km)	Commute time (min)
	2.7	25
	4.1	33
i	1.0	15
	5.2	45
	2.8	22

Linear regression: least squares (1d)

- Let $d = 1$, then:

$$C(w) = \frac{1}{2n}\sum_{i=1}^{n}\left(y^i - x^i w\right)^2$$

What shape does this function have?

	x	y
	Dist (km)	Commute time (min)
	2.7	25
	4.1	33
i	1.0	15
	5.2	45
	2.8	22

Linear regression: least squares (1d)

- Let $d = 1$, then:

$$C(w) = \frac{1}{2n}\sum_{i=1}^{n}\left(y^i - x^i w\right)^2 = \frac{1}{2n}\sum_{i=1}^{n}\left(\underbrace{{x^i}^2 w^2}_{\text{quadratic}} - \underbrace{2y^i x^i w}_{\text{linear}} + \underbrace{{y^i}^2}_{\text{constant}}\right)$$

How can we find the best w?

i	x: Dist (km)	y: Commute time (min)
	2.7	25
	4.1	33
	1.0	15
	5.2	45
	2.8	22

Linear regression: least squares (1d)

- Let $d = 1$, then:

$$C(w) = \frac{1}{2n}\sum_{i=1}^{n}\left(y^i - x^i w\right)^2 = \frac{1}{2n}\sum_{i=1}^{n}\left(\underbrace{{x^i}^2 w^2}_{\text{quadratic}} - \underbrace{2y^i x^i w}_{\text{linear}} + \underbrace{{y^i}^2}_{\text{constant}}\right)$$

- It then follows that:

$$C'(w) = \frac{1}{2n}\sum_{i=1}^{n}\left(2{x^i}^2 w - 2y^i x^i\right) = \frac{1}{n}\sum_{i=1}^{n}{x^i}^2 w - \frac{1}{n}\sum_{i=1}^{n} y^i x^i$$

i	x: Dist (km)	y: Commute time (min)
	2.7	25
	4.1	33
	1.0	15
	5.2	45
	2.8	22

Linear regression: least squares (1d)

- Let $d = 1$, then:

$$C(w) = \frac{1}{2n}\sum_{i=1}^{n}\left(y^i - x^i w\right)^2 = \frac{1}{2n}\sum_{i=1}^{n}\left(\underbrace{{x^i}^2 w^2}_{\text{quadratic}} - \underbrace{2y^i x^i w}_{\text{linear}} + \underbrace{{y^i}^2}_{\text{constant}}\right)$$

- It then follows that:

$$C'(w) = \frac{1}{2n}\sum_{i=1}^{n}\left(2{x^i}^2 w - 2y^i x^i\right) = \frac{1}{n}\sum_{i=1}^{n}{x^i}^2 w - \frac{1}{n}\sum_{i=1}^{n} y^i x^i$$

- According to the first-order optimality condition, we know the optimal w^* satisfies

$$C'(w^*) = 0 \Longrightarrow \frac{1}{n}\sum_{i=1}^{n}{x^i}^2 w^* = \frac{1}{n}\sum_{i=1}^{n} y^i x^i$$

- It then follows that:

$$w^* = \frac{\sum_{i=1}^{n} y^i x^i}{\sum_{i=1}^{n} {x^i}^2}$$

Linear regression: toy example

- Commute time on bus
  - Want to predict commute time to University
  - Input variables (features)?
    - Distance to University
    - Day of the week
  - Output / target?
    - Commute time
  - Data
    - day = 1 if weekday, day = 0 otherwise

Dist (km)	Day	Commute time (min)
2.7	1	25
4.1	1	33
1.0	0	15
5.2	1	45
2.8	0	22

$$f(\boldsymbol{x}) = w_1 x_1 + \cdots + w_d x_d = (w_0 + w_1 + \cdots + w_d)\begin{pmatrix}1\\ \boldsymbol{x}\end{pmatrix} = \boldsymbol{w}^T\boldsymbol{x}$$

Linear regression: matrix form

Dist (km)	Day	Commute time (min)
x_1	x_2	y
2.7	1	25
4.1	1	33
1.0	0	15
5.2	1	45
2.8	0	22

- Let's recall that $C(\boldsymbol{w}) = \frac{1}{2n}\sum_{i=1}^{n}\left(y^i - {\boldsymbol{x}^i}^T\boldsymbol{w}\right)^2$
- Let's consider ${\boldsymbol{x}^i}^T = \left(x_1^i, x_2^i, \dots, x_d^i\right)$

$$X = \begin{pmatrix} {\boldsymbol{x}^1}^T \\ \vdots \\ {\boldsymbol{x}^n}^T \end{pmatrix} \in \mathbb{R}^{n\times d}, \boldsymbol{y} = \begin{pmatrix} y^1 \\ \vdots \\ y^n \end{pmatrix} \in \mathbb{R}^n \Longrightarrow X\boldsymbol{w} - \boldsymbol{y} = \begin{pmatrix} {\boldsymbol{x}^1}^T\boldsymbol{w} - y^1 \\ \vdots \\ {\boldsymbol{x}^n}^T\boldsymbol{w} - y^n \end{pmatrix}$$

$$\boldsymbol{y} = \begin{pmatrix} 25 \\ 33 \\ 15 \\ 45 \\ 22 \end{pmatrix}, X = \begin{pmatrix} & 2.7 & 1 \\ & 4.1 & 1 \\ & 1.0 & 0 \\ & 5.2 & 1 \\ & 2.8 & 0 \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}$$

Linear regression: matrix form

Dist (km)	Day	Commute time (min)
x_1	x_2	y
2.7	1	25
4.1	1	33
1.0	0	15
5.2	1	45
2.8	0	22

- Let's recall that $C(\boldsymbol{w}) = \frac{1}{2n}\sum_{i=1}^{n}\left(y^i - \boldsymbol{x}^{i^T}\boldsymbol{w}\right)^2$
- Let's consider $\boldsymbol{x}^{i^T} = (x_1^i, x_2^i, \dots, x_d^i)$

$$X = \begin{pmatrix} \boldsymbol{x}^{1^T} \\ \vdots \\ \boldsymbol{x}^{n^T} \end{pmatrix} \in \mathbb{R}^{n \times d}, \boldsymbol{y} = \begin{pmatrix} y^1 \\ \vdots \\ y^n \end{pmatrix} \in \mathbb{R}^n \Longrightarrow X\boldsymbol{w} - \boldsymbol{y} = \begin{pmatrix} \boldsymbol{x}^{1^T}\boldsymbol{w} - y^1 \\ \vdots \\ \boldsymbol{x}^{n^T}\boldsymbol{w} - y^n \end{pmatrix}$$

$$\boldsymbol{y} = \begin{pmatrix} 25 \\ 33 \\ 15 \\ 45 \\ 22 \end{pmatrix}, X = \begin{pmatrix} & 2.7 & 1 \\ & 4.1 & 1 \\ & 1.0 & 0 \\ & 5.2 & 1 \\ & 2.8 & 0 \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix}$$

$$\hat{\boldsymbol{y}} = X\boldsymbol{w}$$

$$C(\boldsymbol{w}) = \frac{1}{2n}(X\boldsymbol{w} - \boldsymbol{y})^T(X\boldsymbol{w} - \boldsymbol{y}) = \frac{1}{2n}(\boldsymbol{w}^T X^T - \boldsymbol{y}^T)(X\boldsymbol{w} - \boldsymbol{y})$$
$$= \frac{1}{2n}(\boldsymbol{w}^T X^T X\boldsymbol{w} - 2\boldsymbol{w}^T X^T \boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y})$$

Linear regression: closed-form solution

- Let's recall the objective function:

$$C(\boldsymbol{w}) = \frac{1}{2n}\left(\underbrace{\boldsymbol{w}^T X^T X \boldsymbol{w}}_{\text{quadratic}} - \underbrace{2\boldsymbol{w}^T X^T \boldsymbol{y}}_{\text{linear}} + \underbrace{\boldsymbol{y}^T \boldsymbol{y}}_{\text{constant}}\right)$$

- The gradient of $C(\boldsymbol{w})$ is:

$$\nabla C(\boldsymbol{w}) = \frac{1}{2n}(2X^T X \boldsymbol{w} - 2X^T \boldsymbol{y})$$

- By the first-order optimality condition, we know that optimal $\boldsymbol{w}^*$ satisfies:

$$\nabla C(\boldsymbol{w}^*) = \frac{1}{n}(X^T X \boldsymbol{w}^* - X^T \boldsymbol{y}) = 0 \Longrightarrow X^T X \boldsymbol{w}^* = X^T \boldsymbol{y}$$

- If $X^T X$ is invertible, we get $\boldsymbol{w}^* = (X^T X)^{-1} X^T \boldsymbol{y}$

Linear regression

- Adding a feature for bias term
  - We can add 1 to get an expanded feature vector

Dist (km)	Day	Commute time (min)
x_1	x_2	y
2.7	1	25
4.1	1	33
1.0	0	15
5.2	1	45
2.8	0	22

$\Longrightarrow$

One	Dist (km)	Day	Commute time (min)
x_0	x_1	x_2	y
1	2.7	1	25
1	4.1	1	33
1	1.0	0	15
1	5.2	1	45
1	2.8	0	22

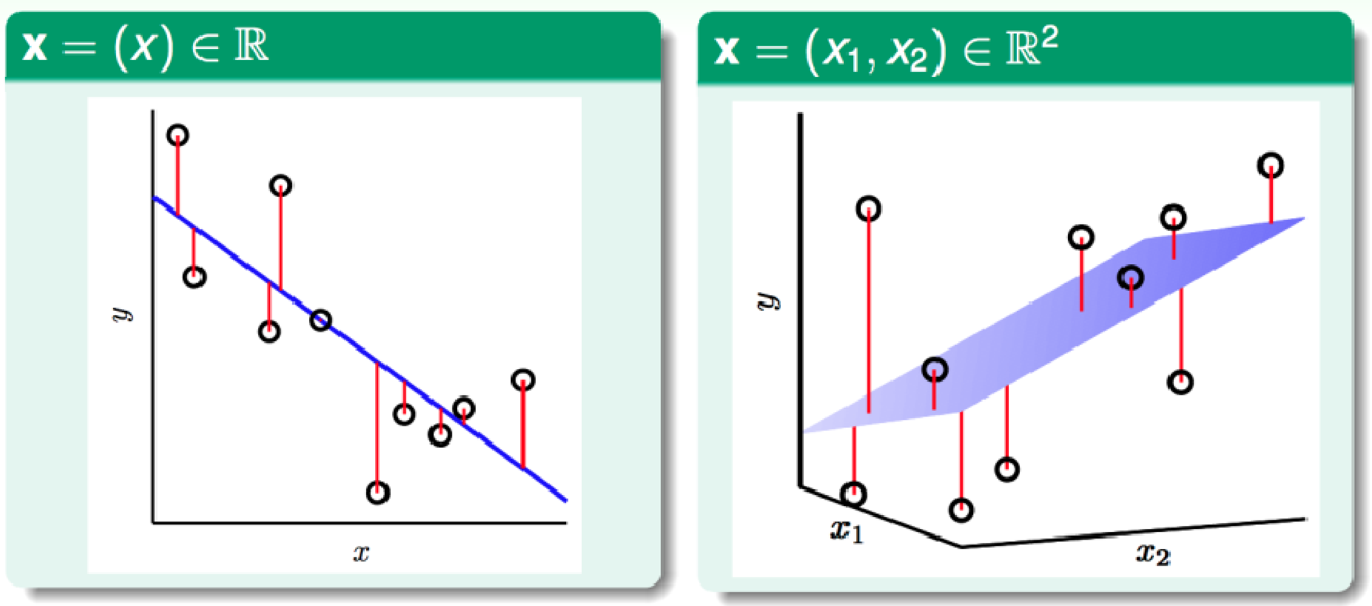
- This allows us to consider the bias in the linear model:

$$f(\boldsymbol{x}) = w_0 + w_1 x_1 + \cdots + w_d x_d = (w_0 + w_1 + \cdots + w_d)\begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix} = \boldsymbol{w}^T \overline{\boldsymbol{x}}$$

- For brevity, we use notation $\boldsymbol{x}$ to represent the extended feature $\overline{\boldsymbol{x}}$:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

Linear regression: illustration



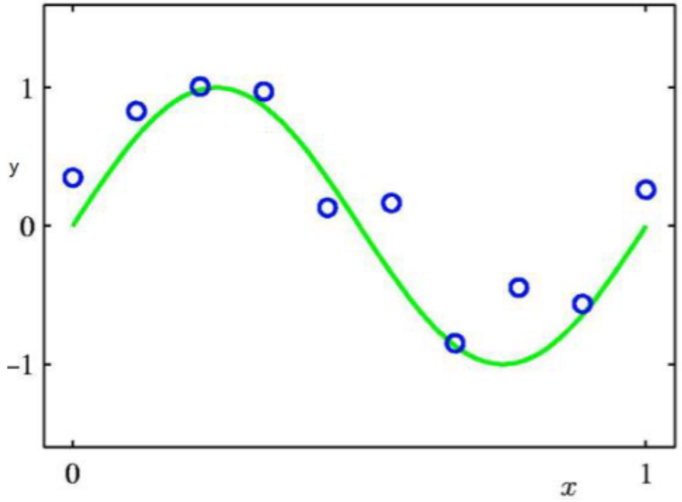
Summary: linear regression

- Linear regression (or least square regression)
  - model linear relationship between input and output (task T)
  - Example points (experience E)
  - mean square error as loss function (performance P)
  - closed-form solution (or exact solution)
  - Add 1s-feature to allow for bias

One	Dist (km)	Day	Commute time (min)
x_0	x_1	x_2	y
1	2.7	1	25
1	4.1	1	33
1	1.0	0	15
1	5.2	1	45
1	2.8	0	22

Polynomial regression

- Suppose we want to model the following data



- The input-output relationship is nonlinear!
- How about we try to fit a polynomial?
    - This is known as <u>polynomial regression</u>

$$f(x) = w_0 + w_1 x + w_1 (x)^2 \ldots, w_M (x)^M$$

where $(x)^i$ denotes i^{th} power of x.

- Do we need to derive a whole new regression algorithm?

Remember the 1-feature?

Dist (km)	Day	Commute time (min)
x_1	x_2	y
2.7	1	25
4.1	1	33
1.0	0	15
5.2	1	45
2.8	0	22

$\Longrightarrow$

One	Dist (km)	Day	Commute time (min)
x_0	x_1	x_2	y
1	2.7	1	25
1	4.1	1	33
1	1.0	0	15
1	5.2	1	45
1	2.8	0	22

Polynomial regression: feature mappings

- Define the feature map:

$$\phi(x) = \begin{pmatrix} 1 \\ x \\ (x)^2 \\ (x)^3 \end{pmatrix}$$

- Polynomial regression model now becomes a linear model w.r.t. the new features

$$f(x) = \boldsymbol{w}^T \phi(x) = w_0 + w_1 x + w_2 (x)^2 + w_3 (x)^3 = \phi(x)^T \boldsymbol{w}$$

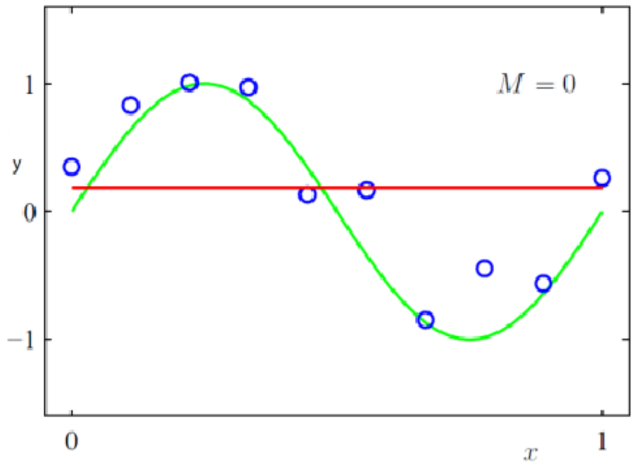
We've transformed a univariate nonlinear problem to a multivariate linear problem!

- The derivations and algorithms so far in this lecture remain the same!

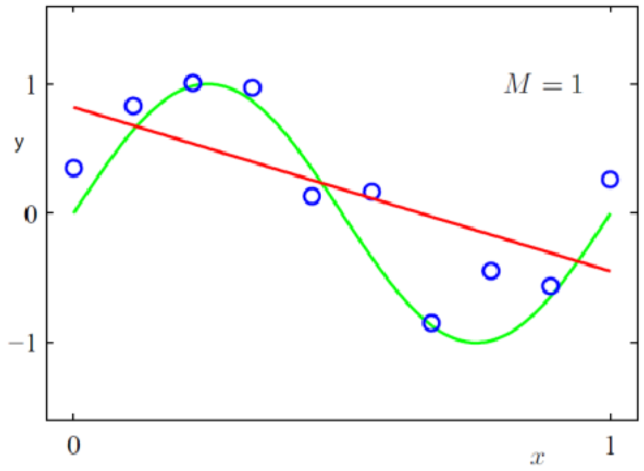
$$X = \begin{pmatrix} \boldsymbol{x}^{1^T} \\ \boldsymbol{x}^{2^T} \\ \vdots \\ \boldsymbol{x}^{n^T} \end{pmatrix} \longmapsto \begin{pmatrix} \phi(x^1)^T \\ \phi(x^2)^T \\ \vdots \\ \phi(x^n)^T \end{pmatrix} = \begin{pmatrix} 1 & x^1 & (x^1)^2 & (x^1)^3 \\ 1 & x^2 & (x^2)^2 & (x^2)^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x^n & (x^n)^2 & (x^n)^3 \end{pmatrix} = \bar{X}$$

Polynomial regression: fitting polynomials

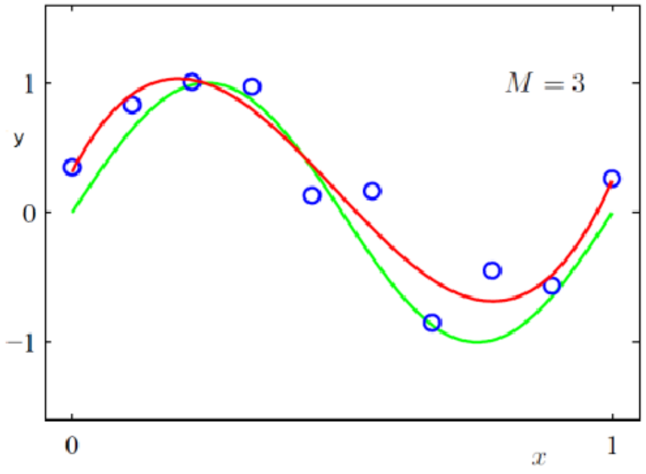
$y = w_0$



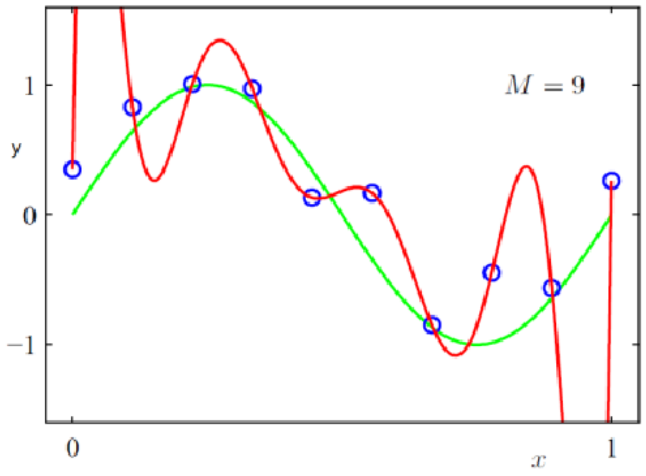
$y = w_0 + w_1 x$



$y = w_0 + w_1 x + w_2 (x)^2 + w_3 (x)^3$



$y = w_0 + w_1 x + w_2 (x)^2 + \cdots + w_9 (x)^9$



Polynomial regression: regularisation

- Regularised least squares regression
  - Given dataset $D = \{(\boldsymbol{x}^1, y^1), (\boldsymbol{x}^2, y^2), \dots, (\boldsymbol{x}^n, y^n)\}$ and a regularisation parameter $\lambda > 0$, find a model to minimise:

$$C(\boldsymbol{w}) = \underbrace{\frac{1}{2n}(\boldsymbol{w}^T X^T X \boldsymbol{w} - 2\boldsymbol{w}^T X^T \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{y})}_{\text{fitting to data}} + \underbrace{\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2}_{\text{regulariser}}$$