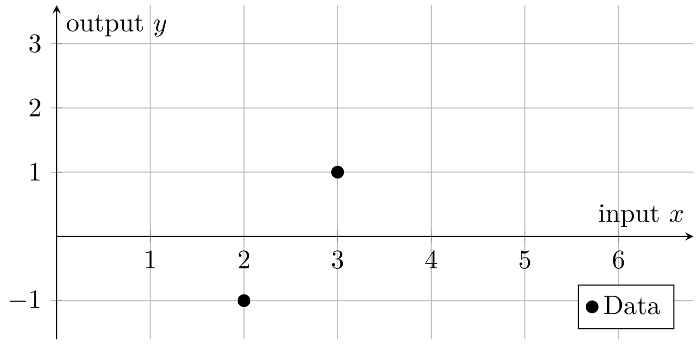
Linear regression

1.1 Linear regression

(a) Assume that you record a scalar input x and a scalar output y. First, you record $x_1 = 2, y_1 = -1$, and thereafter $x_2 = 3, y_2 = 1$. Assume a linear regression model $y = \theta_0 + \theta_1 x + \epsilon$ and learn the parameters with maximum likelihood $\widehat{\boldsymbol{\theta}}$ with the assumption $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Use the model to predict the output for the test input $x_\star = 4$, and add the model to the plot below:



(b) Now, assume you have made a third observation $y_3 = 2$ for $x_3 = 4$ (is that what you predicted in (a)?). Update the parameters $\widehat{\boldsymbol{\theta}}$ to all 3 data samples, add the new model to the plot (together with the new data point) and find the prediction for $x_\star = 5$.

(c) Repeat (b), but this time using a model without intercept term, i.e., $y = \theta_1 x + \epsilon$.

(d) Repeat (b), but this time using Ridge Regression with $\gamma = 1$ instead.

(e) You realize that there are actually *two* output variables in the problem you are studying. In total, you have made the following observations:

sample	input x	first output y_1	second output y_2
(1)	2	-1	0
(2)	3	1	2
(3)	4	2	-1

You want to model this as a linear regression with multidimensional outputs (without regularization), i.e.,

$$y_1 = \theta_{01} + \theta_{11} x + \epsilon \tag{1.1}$$

$$y_2 = \theta_{02} + \theta_{12} x + \epsilon \tag{1.2}$$

By introducing, for the general case of p inputs and q outputs, the matrices

$$\underbrace{\begin{bmatrix} y_{11} & \cdots & y_{1q} \\ y_{21} & \cdots & y_{2q} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nq} \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \theta_{01} & \theta_{02} & \cdots & \theta_{0q} \\ \theta_{11} & \theta_{12} & \cdots & \theta_{1q} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2q} \\ \vdots & \vdots & & \vdots \\ \theta_{p1} & \theta_{p2} & \cdots & \theta_{pq} \end{bmatrix}}_{\boldsymbol{\Theta}} + \boldsymbol{\epsilon}, \tag{1.3}$$

try to make an educated guess how the normal equations can be generalized to the multidimensional output case. (A more thorough derivation is found in problem 1.5). Use your findings to compute the least square solution $\widehat{\boldsymbol{\Theta}}$ to the problem now including both the first output y_1 and the second output y_2.

1.2 Nonlinear transformations of input variables

Assume our data set only contains two variables y and v. What are the transformed inputs/features if we want to learn a linear regression model on the form?

(a) $y = \theta_0 + \theta_1 v + \epsilon$

(b) $y = \theta_0 + \theta_1 v + \theta_2 v^2 + \theta_3 v^3 + \theta_4 v^4 + \epsilon$

(c) $y = \theta_0 + \theta_1 v + \theta_2 \cos(v) + \theta_3 \sin(v) + \epsilon$

(d) $y = \theta_0 + c \sin(v + \phi) + \epsilon$ (neither c nor ϕ are known) *Hint:* $\sin(v + \phi) = \cos(\phi)\sin(v) + \sin(\phi)\cos(v)$.

(e) $y = \theta_0 + \max(\theta_1 v, \theta_2 v^2) + \epsilon$?

1.3 Deriving least squares from maximum likelihood

We assume that each ϵ in the linear regression model is independently distributed as

$$\epsilon \sim \mathcal{N}\left(0,\, \sigma_\epsilon^2\right). \tag{1.4a}$$

We are interested in finding the maximum likelihood estimate of $\boldsymbol{\theta}$

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{y} \,|\, \mathbf{X}; \boldsymbol{\theta}) \qquad \text{where} \qquad p(\mathbf{y} \,|\, \mathbf{X}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \,|\, \mathbf{x}_i; \boldsymbol{\theta}). \tag{1.4b}$$

(The factorization can be done since we assume ϵ to be independent for each data point i.) Show that this implies the least squares problem

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \left(\theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} - y_i\right)^2 \right\}. \tag{1.5}$$

Hint: You only need to show that this implies the least squares problem, not solve the least squares problem (i.e., derive the normal equations). If z is distributed as a Gaussian distribution with mean m and variance σ^2, its probability density is

$$\mathcal{N}\left(z \,|\, m,\, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(z - m\right)^2\right).$$

1.4 The inverse of $\mathbf{X}^\mathsf{T}\mathbf{X}$

(a) Prove that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is a positive definite matrix if and only if the columns of $\mathbf{X}$ are linearly independent.
Hint: A matrix $\mathbf{A}$ is by definition positive definite if and only if $v^\mathsf{T}\mathbf{A}v > 0$ for all column vectors $v \neq 0$.

(b) Use the result from (a) to prove that the unique solution $\widehat{\theta}$ from the normal equations exists if and only if the columns of $\mathbf{X}$ are linearly independent.

(c) Interpret (b) in terms of what properties your data matrix $\mathbf{X}$ needs to have. At least how many data samples do you need to have? What intuitive explanation can you give in the case there is no unique solution $\widehat{\boldsymbol{\theta}}$ to the least squares problem?

1.5 Optimum of multivariable least square

Consider again least squares with multivariable output. With the notation $\mathbf{y}_i = [y_{i1}\, y_{i2} \ldots y_{iq}]^\mathsf{T}$ and $\mathbf{x}_i = [1\, x_{i1}\, x_{i2} \ldots x_{ip}]^\mathsf{T}$. Prove that (1.11) (your finding on problem 1.1(e)) is a stationary point of the multivariable least squares criterion

$$\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\Theta}^\mathsf{T}\mathbf{x}_i)^\mathsf{T}(\mathbf{y}_i - \boldsymbol{\Theta}^\mathsf{T}\mathbf{x}_i), \tag{1.6}$$

if you assume that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is invertible. Three useful equations are

(i) $\frac{\partial}{\partial \mathbf{M}}\left(\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{M}^\mathsf{T}\mathbf{v}\right) = 2\mathbf{v}\mathbf{v}^\mathsf{T}\mathbf{M}$ for any column vector $\mathbf{v}$ of appropriate size.

(ii) $\frac{\partial}{\partial \mathbf{M}}\left(\mathbf{v}^\mathsf{T}\mathbf{M}\mathbf{u}\right) = \mathbf{v}\mathbf{u}^\mathsf{T}$ for any column vectors $\mathbf{u}, \mathbf{v}$ of appropriate sizes.

(iii) $\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\mathsf{T} = \mathbf{X}^\mathsf{T}\mathbf{X}$ and $\sum_{i=1}^{n} \mathbf{x}_i\mathbf{y}_i^\mathsf{T} = \mathbf{X}^\mathsf{T}\mathbf{Y}$.

Solutions

1.1 (a) Write the problem as $\underbrace{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}}_{\mathbf{X}} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}}_{\mathbf{\Theta}} + \boldsymbol{\epsilon}$. The maximum likelihood solution to this problem is equivalent to the least square solution given by $\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\mathbf{\Theta}} = \mathbf{X}^\mathsf{T}\mathbf{y}$. We thus solve it (using, e.g., Gauss elimination),

$$\begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}^\mathsf{T} \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} \widehat{\mathbf{\Theta}} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}^\mathsf{T} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 2 & 5 \\ 5 & 13 \end{bmatrix} \widehat{\mathbf{\Theta}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Rightarrow \widehat{\mathbf{\Theta}} = \begin{bmatrix} -5 \\ 2 \end{bmatrix}. \tag{1.7}$$

The prediction for $x_\star = 4$ becomes $\widehat{y}_\star = \widehat{\theta}_0 + \widehat{\theta}_1 x_\star = 3$.

(b) Again, the solution is given by the normal equations $\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\mathbf{\Theta}} = \mathbf{X}^\mathsf{T}\mathbf{y}$

$$(\mathbf{X}^\mathsf{T}\mathbf{X})\widehat{\mathbf{\Theta}} = \mathbf{X}^\mathsf{T}\mathbf{y} \Rightarrow \left(\begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \right) \widehat{\mathbf{\Theta}} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \Rightarrow \cdots \Rightarrow \widehat{\mathbf{\Theta}} = \frac{1}{6} \begin{bmatrix} -23 \\ 9 \end{bmatrix}. \tag{1.8}$$

The prediction for $x_\star = 5$ is hence $\widehat{y}_\star = \widehat{\theta}_0 + \widehat{\theta}_1 x_\star = \frac{11}{3} \approx 3.67$.

(c) With no intercept term, we get another $\mathbf{X}$ matrix, $\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$, and hence

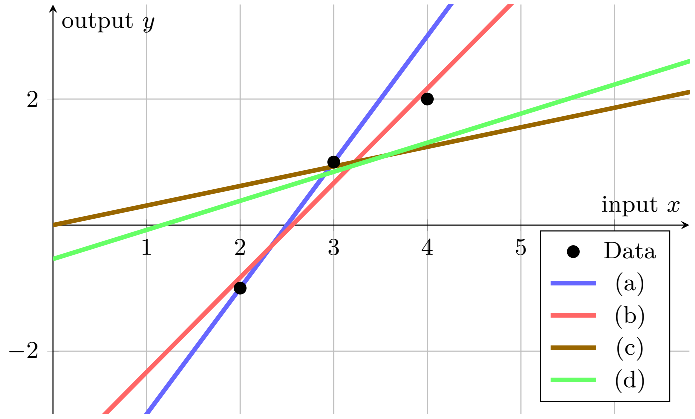
$$\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\mathbf{\Theta}} = \mathbf{X}^\mathsf{T}\mathbf{y} \Rightarrow \begin{bmatrix} 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \widehat{\mathbf{\Theta}} = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \Rightarrow \cdots \Rightarrow \widehat{\mathbf{\Theta}} = \frac{9}{29}, \tag{1.9}$$

with prediction $\widehat{y}_\star = \widehat{\theta}_1 x_\star = \frac{45}{29} \approx 1.55$ for $x_\star = 5$.

(d) We now have to use the solution to the Ridge Regression problem instead, $(\mathbf{X}^\mathsf{T}\mathbf{X} + I_2)\widehat{\boldsymbol{\theta}} = \mathbf{X}^\mathsf{T}\mathbf{y}$,

$$(\mathbf{X}^\mathsf{T}\mathbf{X} + I_2)\widehat{\mathbf{\Theta}} = \mathbf{X}^\mathsf{T}\mathbf{y} \Rightarrow \left(\begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \widehat{\mathbf{\Theta}} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \Rightarrow \cdots \Rightarrow \widehat{\mathbf{\Theta}} = \frac{1}{39} \begin{bmatrix} -21 \\ 18 \end{bmatrix}. \tag{1.10}$$

The prediction for $x_\star = 5$ is hence $\widehat{y}_\star = \widehat{\theta}_0 + \widehat{\theta}_1 x_\star = \frac{69}{39} \approx 1.77$.



(e) The extension of the normal equations are

$$\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\mathbf{\Theta}} = \mathbf{X}^\mathsf{T}\mathbf{Y}. \tag{1.11}$$

Note that this is equivalent to making a separate least square computation for each column in $\mathbf{Y}$.

$$\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\mathbf{\Theta}} = \mathbf{X}^\mathsf{T}\mathbf{Y} \Rightarrow \left(\begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \right) \widehat{\mathbf{\Theta}} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 1 & 2 \\ 2 & -1 \end{bmatrix} \Rightarrow \cdots \Rightarrow \widehat{\mathbf{\Theta}} = \frac{1}{6} \begin{bmatrix} -23 & 11 \\ 9 & -3 \end{bmatrix}. \tag{1.12}$$

Note that the first column is identical to (b).

1.2 (a) No extra inputs are needed, $x_1 = v$.

(b) $x_1 = v, x_2 = v^2, x_3 = v^3, x_4 = v^4$.

(c) $x_1 = v, x_2 = \cos(v), x_3 = \sin(v)$.

(d) We can write the model as $y = \theta_0 + \underbrace{c\cos(\phi)}_{\theta_1}\sin(v) + \underbrace{c\sin(\phi)}_{\theta_2}\cos(v) + \epsilon$, hence $x_1 = \sin(v), x_2 = \cos(v)$.

(e) The model is not possible to write as a linear regression model.

1.3 According to the exercise description, the data distribution has the form

$$\begin{aligned} p(\mathbf{y} \,|\, \mathbf{X}; \boldsymbol{\theta}) &= \prod_{i=1}^{n} p(y_i \,|\, \mathbf{x}_i; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \mathcal{N}\Big(\underbrace{\theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} - y_i}_{\epsilon} \,|\, 0,\, \sigma_\epsilon^2\Big) \\ &\propto \prod_{i=1}^{n} e^{-\frac{1}{2\sigma_\epsilon^2}(\theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} - y_i)^2} \\ &= \exp\left(\sum_{i=1}^{n} -\frac{1}{2\sigma_\epsilon^2}\left(\theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} - y_i\right)^2\right) \end{aligned} \tag{1.13}$$

Since log is a monotonically increasing function, we can write

$$\begin{aligned} \widehat{\theta} &= \arg\max_{\theta} p(\mathbf{y} \,|\, \mathbf{X}; \boldsymbol{\theta}) \\ &= \arg\max_{\theta} \log p(\mathbf{y} \,|\, \mathbf{X}; \boldsymbol{\theta}) \\ &= \arg\max_{\theta} \left\{\sum_{i=1}^{n} -\frac{1}{2\sigma^2}\left(\theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} - y_i\right)^2\right\} \\ &= \arg\min_{\theta} \left\{\sum_{i=1}^{n} \left(\theta_0 + \theta_1 x_{i1} + \cdots + \theta_p x_{ip} - y_i\right)^2\right\} \end{aligned} \tag{1.14}$$

1.4 (a) A matrix $\mathbf{A}$ is by definition positive definite if and only if $\mathbf{v}^\mathsf{T}\mathbf{A}\mathbf{v} > 0$ for all column vectors $v \neq 0$. We can write $\mathbf{v}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{v} = (\mathbf{X}\mathbf{v})^\mathsf{T}(\mathbf{X}\mathbf{v}) = \|\mathbf{X}\mathbf{v}\|^2$. Then $\|\mathbf{X}\mathbf{v}\|^2 = 0 \Leftrightarrow \mathbf{X}\mathbf{v} = 0$, which means that $\mathbf{v}$ is in the null space of $\mathbf{X}$. If $\mathbf{v} \neq 0$, it would mean that $\mathbf{X}$ would have linearly dependent columns, and the claim follows.

(b) The normal equations are only defined if the inverse of $\mathbf{X}^\mathsf{T}\mathbf{X}$ exists, which it does if and only if $\mathbf{X}^\mathsf{T}\mathbf{X}$ is positive definite (by construction, it cannot be negative definite or indefinite). From (a), this is equivalent to $\mathbf{X}$ having linearly independent columns.

(c) A necessary condition for $\mathbf{X}$ having linearly independent columns means that there has to be at least as many data samples n (i.e., rows), as the number of inputs p (i.e., columns). If the columns of $\mathbf{X}$ are not linearly independent, it means that there is no information in the data that can tell the different inputs apart (e.g., if the very same input is used for all recorded data points). *It does, however, **not** mean that a solution does not exist. The least squares problem has in such cases an infinite number of equally good solutions.*

1.5 Differentiating (1.6) yields with respect to $\boldsymbol{\Theta}$

$$\frac{\partial}{\partial\boldsymbol{\Theta}}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\Theta}^\mathsf{T}\mathbf{x}_i)^\mathsf{T}(\mathbf{y}_i - \boldsymbol{\Theta}^\mathsf{T}\mathbf{x}_i) = \frac{\partial}{\partial\boldsymbol{\Theta}}\sum_{i=1}^{n}\left(\mathbf{y}_i^\mathsf{T}\mathbf{y}_i - 2\mathbf{x}_i^\mathsf{T}\boldsymbol{\Theta}\mathbf{y}_i + \mathbf{x}_i^\mathsf{T}\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}\mathbf{x}_i\right) = /\text{(i) \& (ii)}/ =$$

$$\sum_{i=1}^{n}\left(0 - 2\mathbf{x}_i\mathbf{y}_i^\mathsf{T} + 2\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\boldsymbol{\Theta}\right) = 2\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\mathsf{T}\right)\boldsymbol{\Theta} - 2\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}_i^\mathsf{T}\right) = /\text{(iii)}/ = 2(\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\Theta} - \mathbf{X}^\mathsf{T}\mathbf{Y}) \tag{1.15}$$

Inserting $\widehat{\boldsymbol{\Theta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}$ (since we assumed $\mathbf{X}^\mathsf{T}\mathbf{X}$ to be invertible) into (1.15) gives

$$2(\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\Theta}} - \mathbf{X}^\mathsf{T}\mathbf{Y}) = 2\left((\mathbf{X}^\mathsf{T}\mathbf{X})(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} - \mathbf{X}^\mathsf{T}\mathbf{Y}\right) = \mathbf{0}. \tag{1.16}$$