Department of Electrical Engineering
Machine learning for signal processing
Q3, 2020-2021

5LSL0 - Machine learning for signal processing

Assignment 2 : Non Linear Models

C. Schmitt	1619020
S. Maulod	1457284

Coordinators:
Dr. ir. R. Vullings
Dr. ir. R.J.G. van Slouns

Eindhoven, May 19, 2021

Exercise 1.1 Linear Functions

Q1 We define the following matrix notation for this question.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{x} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \qquad \theta = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{pmatrix}$$

where $w_p = b$ and $x_{n,p} = 1$.
The Gaussian distribution of prediction errors with mean $f(x_i;\theta)$, variance = 1 and unity covariance is given by : $p(y; f(x_i;\theta), \sigma\mathbf{I}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - f(x_i;\theta))^2}{2\sigma}}$

$$\begin{aligned} J(\mathbf{x}, y; \theta) &= -\log \prod_{i=0}^{m-1} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - f(x_i;\theta))^T (y_i - f(x_i;\theta))}{2}} \\ &= \log \frac{m}{\sqrt{2\pi}} + \frac{1}{2} \sum_{i=0}^{m-1} (y_i - f(\mathbf{x_i};\theta))^T (y_i - f(\mathbf{x_i};\theta)) \\ J'(\mathbf{x}, y; \theta) &= \frac{1}{2}(y - \mathbf{x}\theta)^T (y - \mathbf{x}\theta) \end{aligned}$$

We can neglect the first term as it does not contain the parameters w and b, resulting in the cost function $J'(\mathbf{x}, y; \theta)$

Q2 For optimal parameters θ_{opt},

$$\begin{aligned} \theta_{\text{opt}} &= \arg\min_{\theta} J(\mathbf{x}, y; \theta) \\ \frac{\partial}{\partial\theta} J(\mathbf{x}, y; \theta) &= 0 \\ \frac{\partial}{\partial\theta} \frac{1}{2}(y - \mathbf{x}\theta)^T (y - \mathbf{x}\theta) &= 0 \\ -\mathbf{x}^T (y - \mathbf{x}\theta_{\text{opt}}) &= 0 \\ \mathbf{x}^T \mathbf{x}\theta_{\text{opt}} &= \mathbf{x}^T y \\ \theta_{\text{opt}} &= (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T y \end{aligned}$$

Q3 Optimal $\mathbf{w} = \begin{pmatrix} 0.1 \\ 0.4 \end{pmatrix}$, b = 0. The mean squared error of the model is 2.78×10^{-32}, which is very small, thus the process is well described by the regression model.

Q4 Optimal $\mathbf{w} = \begin{pmatrix} 0.1011 \\ 0.4107 \end{pmatrix}$, b = -0.0502. The mean squared error now becomes 1×10^{-4} which is several orders of magnitude worse than in Q3. One way to obtain better estimates would be to increase the number of available data points by taking mores samples.

Q5 The Gaussian distribution of prediction errors with mean $f(x_i;\theta)$, variance = 1 and non-unity covariance is given by : $p(y; f(x_i;\theta), \sigma\mathbf{I}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - f(x_i;\theta))^2}{2\sigma}}$

$$\begin{aligned}
J(\mathbf{x}, y; \theta, \Sigma) &= -\log \prod_{i=0}^{m-1} \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{(y_i - f(x_i;\theta))^T \Sigma^{-1} (y_i - f(x_i;\theta)}{2}} \\
&= \sum_{i=0}^{m-1} \left(-\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{y_i} - \mathbf{f}(\mathbf{x_i};\theta))^{\mathbf{T}} \boldsymbol{\Sigma}^{\mathbf{-1}} (\mathbf{y_i} - \mathbf{f}(\mathbf{x_i};\theta)) \right) \\
J'(\mathbf{x}, y; \theta) &= \frac{1}{2}(y - \mathbf{x}\theta)^T \Sigma^{-1} (y - \mathbf{x}\theta) \\
J'(\mathbf{x}, y; \theta) &= \frac{1}{2}(y - \mathbf{x}\theta)^T \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n} \end{bmatrix} (y - \mathbf{x}\theta)
\end{aligned}$$

We can neglect the first term as it does not contain the parameters w and b, resulting in the cost function $J'(\mathbf{x}, y; \theta)$

The difference between predicted and goal value is weighted by $\frac{1}{\sigma_n}$, meaning that dimensions with low variance are given higher penalties in the cost function

Q6 Optimal $\mathbf{w} = \begin{pmatrix} -0.5 \\ -1 \end{pmatrix}$, b = 1.5

Exercise 1.2 Non linear Functions

Q7 The solutions for the compact expressions are given below

- **ReLU** :

$$\frac{\partial f(x)}{\partial x} = \begin{cases} 1, & \text{if } x > 0 \\ \text{undefined}, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$$

- **Sigmoid** :

$$\begin{aligned}
\frac{\partial f(x)}{\partial x} &= \frac{\partial}{\partial x} \frac{1}{e^{-x} + 1} \\
&= -\frac{\frac{\partial}{\partial x}[e^{-x} + 1]}{(e^{-x} + 1)^2} \\
&= \frac{e^{-x}}{(e^{-x} + 1)^2} \\
&= \frac{1}{(e^{-x} + 1)} \frac{e^{-x}}{(e^{-x} + 1)} \\
&= \frac{1}{(e^{-x} + 1)} \frac{1 - 1 + e^{-x}}{(e^{-x} + 1)} \\
&= \frac{1}{(e^{-x} + 1)} \left(\frac{e^{-x} + 1}{(e^{-x} + 1)} - \frac{1}{(e^{-x} + 1)} \right) \\
&= \sigma(x)\,(1 - \sigma(x))
\end{aligned}$$

- **Softmax**:
  We first calculate the elementwise derivative, x_a for the Jacobian matrix of the Softmax

function.

$$\frac{\partial f(\mathbf{x})_j}{\partial x_a} = \frac{\partial}{\partial x_a} \frac{\exp(x_j)}{\sum_{i=0}^{K-1} \exp(x_i)}$$

Using the quotient rule, if $f(x) = \frac{g(x)}{h(x)}$ then $f'(x) = \frac{g'(x)h(x)-g(x)h'(x)}{h(x)^2}$
$g_j = \exp(x_j)$ and $h_j = \sum_{i=0}^{K-1} \exp(x_i)$
$\frac{\partial g_j}{\partial x_a} = \begin{cases} \exp(x_a), & \text{if } j = a \\ 0, & \text{otherwise} \end{cases}$
$\frac{\partial h_j}{\partial x_a} = \exp(x_a)$
When j=a,

$$\begin{aligned} \frac{\partial}{\partial x_a} \frac{\exp(x_j)}{\sum_{i=0}^{K-1} \exp(x_i)} &= \frac{\exp(x_a) \sum_{i=0}^{K-1} \exp(x_i) - \exp(x_j)\exp(x_a)}{(\sum_{i=0}^{K-1} \exp(x_i))^2} \\ &= \frac{\exp(x_a)}{\sum_{i=0}^{K-1} \exp(x_i)} \cdot \frac{\sum_{i=0}^{K-1} \exp(x_i) - \exp(x_j)}{\sum_{i=0}^{K-1} \exp(x_i)} \\ &= \frac{\exp(x_a)}{\sum_{i=0}^{K-1} \exp(x_i)} \left(1 - \frac{\exp(x_j)}{\sum_{i=0}^{K-1} \exp(x_i)}\right) \\ &= f(\mathbf{x})_a(1 - f(\mathbf{x})_j) \end{aligned}$$

when $j \neq a$

$$\begin{aligned} \frac{\partial}{\partial x_a} \frac{\exp(x_j)}{\sum_{i=0}^{K-1} \exp(x_i)} &= \frac{0 - \exp(x_j)\exp(x_a)}{(\sum_{i=0}^{K-1} \exp(x_i))^2} \\ &= -f(\mathbf{x})_a f(\mathbf{x})_j \end{aligned}$$

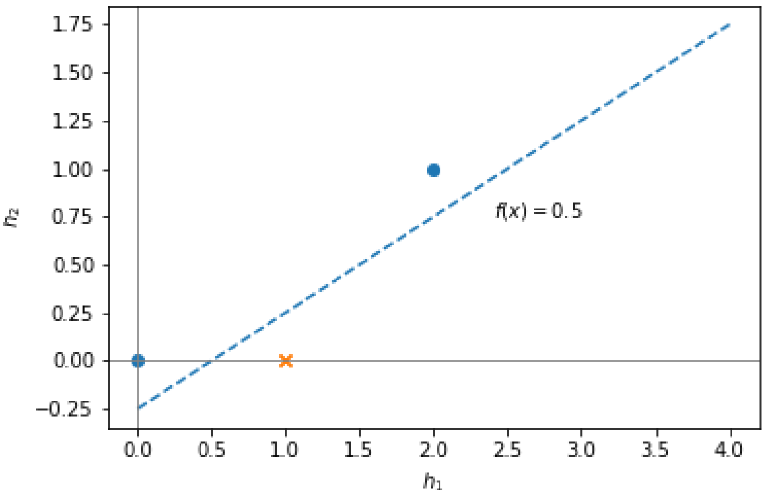
$$\frac{\partial f(x)_j}{\partial x_a} = \begin{cases} f(\mathbf{x})_a(1 - f(\mathbf{x})_j), & \text{if } j = a \\ -f(\mathbf{x})_a f(\mathbf{x})_j, & \text{if } j \neq a \end{cases}$$

Q8 **ReLU** : Gradient keeps increasing linearly
Sigmoid : Gradient becomes 0
Softmax : Gradient is bounded between -1 and 1 regardless of x

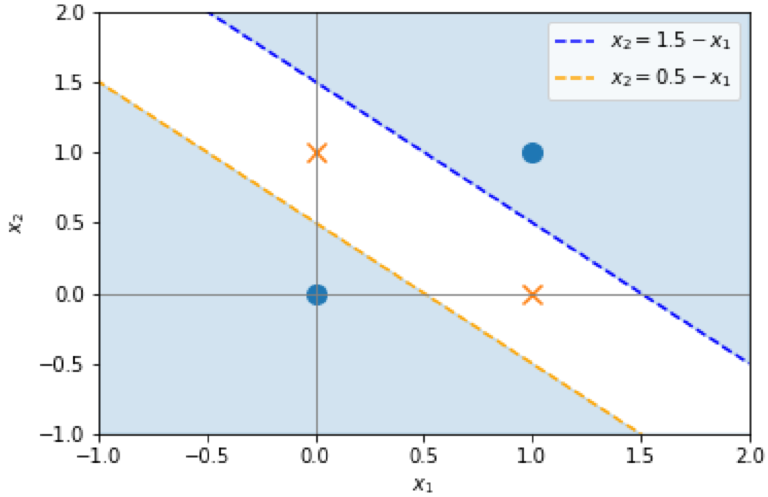
Exercise 1.3 Shallow nonlinear models

Q9 When there is a linear model, this results in a very poor model as seen in Q6. If the mapping was linear, then the two functions and the mapping could be reduced into one linear function, the new linear model would have the same problems as using one linear function

Q10



Q11 We see that there are two decision boundaries in the input space reconstruction: Of the four possible combinations of h_1 and h_2 being bigger than 0 or equal to 0 only two have valid solutions. For f(x)=0.5, the decision boundaries are $x_2 < 0.5 - x_1$ (for $h_2 = 0$, $h_1 > 0$) and $x_2 > 1.5 - x_1$ (for $h_1 = 0$, $h_2 = 0$).



Exercise 1.4 Binary classification with logistic regression

Q12 Softmax: not only does it ensure outputs are between 0 and 1, it also ensures all outputs add up to 1 - these can therefore be interpreted like probabilities

Q13

$$\begin{aligned}
\frac{\partial J}{\partial p} &= -\frac{\partial}{\partial p}\sum_{i=0}^{m-1}\Big(y_i \text{log}(p^{(i)}) + (1-y^{(i)})\text{log}(1-p^{(i)})\Big) \\
&= -\frac{\partial}{\partial p}\sum_{i=0}^{m-1}\left(\frac{y^{(i)}}{p^{(i)}} + \frac{1-y^{(i)}}{1-p^{(i)}})(-1)\right) \\
&= -\sum_{i=0}^{m-1}\left(\frac{y^{(i)}}{p^{(i)}} - \frac{1-y^{(i)}}{1-p^{(i)}}\right)
\end{aligned}$$

Q14

$$\begin{aligned}
\frac{\partial p}{\partial f} &= \frac{\sigma(f(x^{(i)};w)}{\partial f} \\
&= \frac{\partial}{\partial f}\frac{1}{1+\exp(-f(x^{(i)};w))} \\
&= \frac{\exp(-f(x^{(i)};w))}{(\exp(-f(x^{(i)};w))+1)^2} \\
&= \sigma(f(x^{(i)};w))\Big(1-\sigma(f(x^{(i)};w))\Big)
\end{aligned}$$

Q15

$$\begin{aligned}
\frac{\partial f(x^{(i)};w)}{\partial w} &= \frac{\partial}{\partial w}w^T x \\
&= x
\end{aligned}$$

This corresponds to the gradient descent step of the linear adaptive filter.

Q16

$$\begin{aligned}
\frac{\partial J}{\partial w} &= -\frac{\partial J}{\partial p}\frac{\partial p}{\partial f}\frac{\partial f}{\partial w} \\
&= -\sum_{i=0}^{m-1}\left(\frac{y^{(i)}}{p^{(i)}} - \frac{1-y^{(i)}}{1-p^{(i)}}\right)\cdot \sigma(f(x^{(i)};w))\Big(1-\sigma(f(x^{(i)};w))\Big)\cdot x \\
&= -\sum_{i=0}^{m-1}\left(\frac{y^{(i)}}{\sigma(f(x^{(i)};w))} - \frac{1-y^{(i)}}{1-\sigma(f(x^{(i)};w))}\right)\cdot \sigma(f(x^{(i)};w))\Big(1-\sigma(f(x^{(i)};w))\Big)\cdot x \\
&= -\sum_{i=0}^{m-1}x_i\Big(y_i(1-\sigma(f(x^{(i)};w)) - (1-y_i)(\sigma(f(x^{(i)};w))\Big) \\
&= -\sum_{i=0}^{m-1}x_i\Big(y_i - \sigma(f(x^{(i)};w))\Big)
\end{aligned}$$

Q17 $\frac{\partial J}{\partial W_2}$: option B

$$\begin{aligned}\frac{\partial J}{\partial w^{(2)}} &= \frac{\partial J}{\partial p}\frac{\partial p}{\partial f^{(2)}}\frac{\partial f^{(2)}}{\partial w^{(2)}}\\ &= -\sum_{i=0}^{m-1}\left(\frac{y^{(i)}}{p^{(i)}}-\frac{1-y^{(i)}}{1-p^{(i)}}\right)\sigma(x)\left(1-\sigma(x)\right)\frac{\partial f^{(2)}}{\partial w^{(2)}}\\ &= -\sum_{i=0}^{m-1}\left(y^{(i)}-\sigma(f^{(2)})\right)\frac{\partial f^{(2)}}{\partial w^{(2)}}\\ &= \begin{cases}-\sum_{i=0}^{m-1}\left(y^{(i)}-\sigma(f^{(2)})\right)(W^{(1)}X+b^{(1)}), & \text{if } W^{(1)}x^{(i)}+b^{(i)}\geq 0\\ 0, & \text{else}\end{cases}\end{aligned}$$

$\frac{\partial J}{\partial w^{(1)}}$: option A

$$\begin{aligned}\frac{\partial J}{\partial w^{(1)}} &= \frac{\partial J}{\partial p}\frac{\partial p}{\partial f^{(2)}}\frac{\partial f^{(2)}}{\partial f^{(1)}}\frac{\partial f_1}{\partial w^{(1)}}\\ &= -\sum_{i=0}^{m-1}\left(y^{(i)}-\sigma(f^{(2)})\right)\frac{\partial f^{(2)}}{\partial f^{(1)}}\frac{\partial f^{(1)}}{\partial w^{(1)}}\\ &= \begin{cases}-\sum_{i=0}^{m-1}\left(y^{(i)}-\sigma(f^{(2)})\right)(w^{(2)})^T\frac{\partial(W^{(1)}x^{(i)}+b_1)}{\partial w^{(1)}}, & \text{if } W^{(1)}x^{(i)}+b^{(i)}\geq 0\\ 0, & \text{else}\end{cases}\\ &= \begin{cases}-\sum_{i=0}^{m-1}\left(y^{(i)}-\sigma(f^{(2)})\right)(w^{(2)})^T x^{(i)}, & \text{if } W^{(1)}x^{(i)}+b^{(1)}\geq 0\\ 0, & \text{else}\end{cases}\end{aligned}$$

$\frac{\partial J}{\partial b^{(2)}}$: option C

$$\begin{aligned}\frac{\partial J}{\partial b^{(2)}} &= \frac{\partial J}{\partial p}\frac{\partial p}{\partial f^{(2)}}\frac{\partial f^{(2)}}{\partial b^{(2)}}\\ &= -\sum_{i=0}^{m-1}\left(y^{(i)}-\sigma(f^{(2)})\right)\frac{\partial f^{(2)}}{\partial b^{(2)}}\\ &= -\sum_{i=0}^{m-1}\left(y^{(i)}-\sigma(f^{(2)})\right)\end{aligned}$$

$\frac{\partial J}{\partial b_1}$: option B

$$\begin{aligned}\frac{\partial J}{\partial w^{(2)}} &= \frac{\partial J}{\partial p}\frac{\partial p}{\partial f^{(2)}}\frac{\partial f^{(2)}}{\partial b^{(1)}}\\ &= -\sum_{i=0}^{m-1}\left(\frac{y^{(i)}}{p^{(i)}}-\frac{1-y^{(i)}}{1-p^{(i)}}\right)\sigma(x)\left(1-\sigma(x)\right)\frac{\partial f^{(2)}}{\partial b^{(1)}}\\ &= -\sum_{i=0}^{m-1}\left(y^{(i)}-\sigma(f^{(2)})\right)\frac{\partial f^{(2)}}{\partial b^{(1)}}\\ &= \begin{cases}-\sum_{i=0}^{m-1}\left(y^{(i)}-\sigma(f^{(2)})\right)W^{(2)T}, & \text{if } W^{(1)}x^{(i)}+b^{(1)}\geq 0\\ 0, & \text{else}\end{cases}\end{aligned}$$