AM41AN Coursework 2023

Due by Friday the 28 April 2023 at 12:00 (electronic submission only).

Instructions

In the present coursework you are required to perform a number of tasks. All your findings must be presented in the form of a scientific report. Such a report must be written in LaTeX, and it must contain an abstract, introduction, methods, results and conclusions. All formulas must be typed (not scanned) and all your graphs and plots must be original. You are expected to work individually towards the completion of these tasks.

Each section must agree with the following descriptors:

- Abstract: It must summarize the nature of the problem, the methods used to solve it, the results obtained and the conclusions drawn from your results. **5 marks**.
- Introduction: It must contain a mathematical definition of the problem to be solved. Do not be afraid to use mathematics to properly define the problem (in this case, regression). Do not scan your formulas from somewhere else, type them yourself. Do not describe the particular instance you are solving (in the present case your data set has two-dimensional inputs, this is a particular feature of the present instance, the generic regression problem has inputs $\boldsymbol{x}$ in the space $\mathscr{X}$ which is a subset of $\mathbb{R}^d$, where d is undetermined). **20 marks**.
- Methods: Thorough mathematical derivation and description of the method you will apply to solve the problem. You must describe your methods using mathematics. Word descriptions should appear, if so, after the math has done the talking. **20 marks**.
- Results: You present here your findings in the form of plots and tables. Axes must be named and both figures and tables must be captioned. Failure to do so negatively reflects in the presentation component of the assessment. **30 marks**.
- Conclusions: They must be extracted from the results presented in the section above. Conclusions are not a matter of opinion ("I think this could be seen as...") they are matter of fact ("The results show ..."). **15 marks**.
- Presentation: Maximum of **10 marks**. Deviation from the rules presented above will result in mark deductions. Please do not forget to put your name and ID number at the top of your report.

The description of your tasks is as follows:

Download the file `data.dat` from the AM41AN web page in BlackBoard. This file contains the data set $\mathcal{D} = \{(|\boldsymbol{x}_\ell\rangle, t_\ell)\}_{\ell=1}^{L}$ with $\boldsymbol{x}_\ell \in \mathscr{X} \subset \mathbb{R}^2$, $t_\ell \in \mathscr{T} \subset \mathbb{R}$ and $L = 10^4$. The first two numbers per row are the coordinates of the vector $|\boldsymbol{x}_\ell\rangle$, the independent variable, the third is t_n, the target or dependent variable. Some level of noise has been added to the targets.

1. Select at random a subset $\mathcal{D}_t \subset \mathcal{D}$ of size $|\mathcal{D}_t| = 10^3$. This will be your training set.
2. For all vectors $\boldsymbol{x} \in \mathbb{R}^2$ let us define the vector $\boldsymbol{r} \in \mathbb{R}^6$ as $\langle \boldsymbol{r}| := \left(1, x_1, x_2, x_1x_2, x_1^2, x_2^2\right)$. Given the Polynomial of order 2 in $\boldsymbol{x} \in \mathbb{R}^2$ with coefficients $\langle \boldsymbol{\Phi}| = (\phi_0, \ldots, \phi_5)$, i.e.:

$$\langle \boldsymbol{r} | \boldsymbol{\Phi} \rangle = \phi_0 + \phi_1 x_1 + \phi_2 x_2 + \phi_3 x_1 x_2 + \phi_4 x_1^2 + \phi_5 x_2^2 \tag{1}$$

and the loss function defined as:

$$E(\boldsymbol{\Phi}) = \frac{1}{|\mathcal{D}_t|} \sum_{(\boldsymbol{x},t)\in\mathcal{D}_t} \mathcal{E}(\boldsymbol{\Phi}|\boldsymbol{x}, t) \tag{2}$$

$$\mathcal{E}_\ell(\boldsymbol{\Phi}) := \mathcal{E}(\boldsymbol{\Phi}|\boldsymbol{x}_\ell, t_\ell) = \frac{1}{2}\left(t_\ell - \langle \boldsymbol{r}_\ell | \boldsymbol{\Phi} \rangle\right)^2, \tag{3}$$

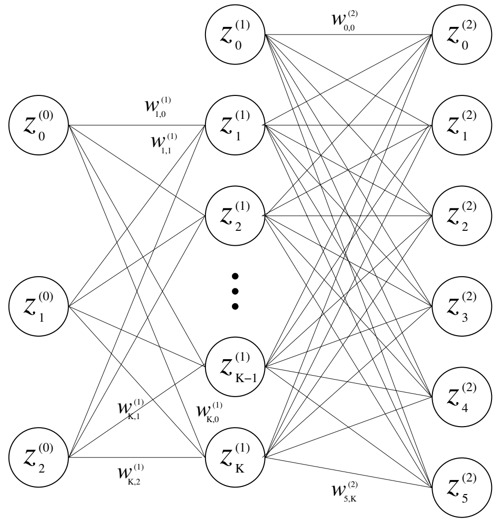

Figure 1: Architecture of the proposed network. The units in the input layer correspond to $z_0^{(0)} = 1$, $z_1^{(0)} = x_1$ and $z_2^{(0)} = x_2$ The $K+1$ units in the hidden layer are $z_0^{(1)} = 1$, and $z_k^{(1)} = \tanh\left(w_{k,0}^{(1)} z_0^{(0)} + w_{k,1}^{(1)} z_1^{(0)} + w_{k,2}^{(1)} z_2^{(0)}\right)$ for all $0 < k \leq K$. The output units are linear, i.e. $z_i^{(2)} = w_{i,0}^{(2)} z_0^{(1)} + \sum_{k=1}^{M} w_{i,k}^{(2)} z_k^{(1)}$ with $0 \leq i \leq 5$.

train by Error Back Propagation (EBP) the network presented in figure 1. This network consists of three layers (input, hidden and output). The units in the input layer correspond to $z_0^{(0)} = 1$, $z_1^{(0)} = x_1$ and $z_2^{(0)} = x_2$. The $K+1$ ($K = 20$) units in the hidden layer are $z_0^{(1)} = 1$, and $z_k^{(1)} = \tanh\left(w_{k,0}^{(1)} + w_{k,1}^{(1)} z_1^{(0)} + w_{k,2}^{(1)} z_2^{(0)}\right)$ for all $0 < k \leq K$. The six output units are linear, i.e. $z_i^{(2)} = w_{i,0}^{(2)} z_0^{(1)} + \sum_{k=1}^{K} w_{i,k}^{(2)} z_k^{(1)}$ with $0 \leq i \leq 5$. Thus, the network performs the following function: $\mathbf{\Phi} : \{1\} \times \mathscr{X} \to \mathbb{R}^6$. The network has to be trained to output the coefficients $\mathbf{\Phi}$ of the polynomial of order 2 $\langle \boldsymbol{r} \,|\mathbf{\Phi} \rangle$ that best fits the data set, i.e. $\left\langle \boldsymbol{r} \,\middle|\boldsymbol{z}^{(2)} \right\rangle$, where $\mathbf{\Phi}(\boldsymbol{x}) = \boldsymbol{z}^{(2)}$.
Hint: The gradient of the loss function has the following components:

$$\frac{\partial}{\partial w_{i,k}^{(2)}} \mathcal{E}(\boldsymbol{z}^{(2)} | \boldsymbol{x}_\ell, t_\ell) = \left(\left\langle \boldsymbol{r}_\ell \,\middle|\boldsymbol{z}^{(2)} \right\rangle - t_\ell \right) r_{i,\ell} z_k^{(1)}, \tag{4}$$

$$\frac{\partial}{\partial w_{k,n}^{(1)}} \mathcal{E}(\boldsymbol{z}^{(2)} | \boldsymbol{x}_\ell, t_\ell) = \left(\left\langle \boldsymbol{r}_\ell \,\middle|\boldsymbol{z}^{(2)} \right\rangle - t_\ell \right) \left[1 - \left(z_k^{(1)} \right)^2 \right] \sum_{i=0}^{5} r_{i,\ell} w_{i,k}^{(2)} \, z_n^{(0)}. \tag{5}$$

The suggested EBP algorithm is:

$$\left[w_{i,k}^{(2)} \right]_{\ell+1} = \left[w_{i,k}^{(2)} \right]_{\ell} - \frac{\eta}{\sqrt{\ell}} \frac{\partial}{\partial w_{i,k}^{(2)}} \mathcal{E}(\boldsymbol{z}^{(2)} | \boldsymbol{x}_\ell, t_\ell), \tag{6}$$

$$\left[w_{k,n}^{(1)} \right]_{\ell+1} = \left[w_{k,n}^{(1)} \right]_{\ell+1} - \frac{\eta}{\sqrt{\ell}} \frac{\partial}{\partial w_{k,n}^{(1)}} \mathcal{E}(\boldsymbol{z}^{(2)} | \boldsymbol{x}_\ell, t_\ell), \tag{7}$$

with a $\eta \approx 2 \times 10^{-3}$ and an initial condition for the weights drawn from a Gaussian distribution centered at zero and with unit standard deviation of 0.1.

3. Use the complementary set $\mathcal{D}_v = \mathcal{D} \backslash \mathcal{D}_t$ to validate your results. To do so separate $\mathcal{D}_v$ into ten subsets $\mathcal{D}_v^{(j)}$, $1 \leq j \leq 10$, such that $\mathcal{D}_v = \bigcup_{j=1}^{10} \mathcal{D}_v^{(j)}$ and $\mathcal{D}_v^{(i)} \bigcap \mathcal{D}_v^{(j)} = \emptyset$ for all $i \neq j$. Then proceed to compute the experimental mean and variance per set of the estimates, keeping the weights from your last update:

$$\hat{\mathbf{\Phi}}^{(j)} := \frac{1}{|\mathcal{D}_v^{(j)}|} \sum_{(\boldsymbol{x},t) \in \mathcal{D}_v^{(j)}} \mathbf{\Phi}(\boldsymbol{x}) \tag{8}$$

$$V[\Phi_i^{(j)}] := \frac{1}{|\mathcal{D}_v^{(j)}|} \sum_{(\boldsymbol{x},t) \in \mathcal{D}_v^{(j)}} \left([\mathbf{\Phi}(\boldsymbol{x})]_i - \left[\hat{\mathbf{\Phi}}^{(j)} \right]_i \right)^2, , \tag{9}$$

where for all $\boldsymbol{v} \in \mathbb{R}^6$, $[\boldsymbol{v}]_i = v_i$ is the i-th component of the vector $\boldsymbol{v}$. With these quantities you can compute the final mean and variance as:

$$\overline{\boldsymbol{\Phi}} := \frac{1}{10} \sum_{j=1}^{10} \hat{\boldsymbol{\Phi}}^{(j)}, \tag{10}$$

$$\overline{V}(\phi_i) := \frac{1}{90} \sum_{j=1}^{10} V[\Phi_i^{(j)}]. \tag{11}$$

Report your final result as the estimate $\overline{\phi}_i$ and its standard deviation $\sqrt{\overline{V}(\phi_i)}$.

4. Make a plot of your estimate $\left\langle \boldsymbol{r} \,\middle|\, \overline{\boldsymbol{\Phi}} \right\rangle$ against $\boldsymbol{x}$, toghether with the scattered points from the training set.
5. Once you have solved the regression task set above, consider the following regularized error function:

$$\tilde{E}\left(|\boldsymbol{w}\rangle\right) = E\left(|\boldsymbol{w}\rangle\right) + \frac{\nu}{2} \left\langle \boldsymbol{w} \,|\boldsymbol{w} \right\rangle, \tag{12}$$

where the vector $|\boldsymbol{w}\rangle$ considered is the last estimate obtained from the implementation of the EBP algorithm and $\nu > 0$ is a suitable parameter. Demonstrate that the Hessian matrix:

$$[\boldsymbol{H}]_{i,j} := \frac{\partial^2 E\left(\mathbf{w}\right)}{\partial w_j \partial w_k} = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},t)\in\mathcal{D}} \frac{\partial^2 \mathcal{E}(\hat{\boldsymbol{\Phi}}|\boldsymbol{x},t)}{\partial w_j \partial w_k},$$

has the following components:

$$\frac{\partial^2 \mathcal{E}_\ell}{\partial w_{i,k}^{(2)} \partial w_{i',k'}^{(2)}} = r_{i,\ell} r_{i',\ell} z_k^{(1)} z_{k'}^{(1)}, \tag{13}$$

$$\frac{\partial^2 \mathcal{E}_\ell}{\partial w_{k',n}^{(1)} \partial w_{i,k}^{(2)}} = r_{i,\ell} \left[1 - \left(z_{k'}^{(1)}\right)^2\right] z_n^{(0)} \left\{ z_k^{(1)} \sum_{i'=0}^{5} r_{i',\ell} w_{i',k'}^{(2)} + \delta_{k,k'} \left[\left\langle \boldsymbol{r}_\ell \,\middle|\, \hat{\boldsymbol{\Phi}} \right\rangle - t_\ell\right] \right\}, \tag{14}$$

$$\frac{\partial^2 \mathcal{E}_\ell}{\partial w_{k',n'}^{(1)} \partial w_{k,n}^{(1)}} = \left[1 - \left(z_k^{(1)}\right)^2\right] z_n^{(0)} z_{n'}^{(0)} \sum_{i'=0}^{5} r_{i',\ell} w_{i',k}^{(2)} \left\{ \left[1 - \left(z_{k'}^{(1)}\right)^2\right] \sum_{i'=0}^{5} r_{i',\ell} w_{i',k'}^{(2)} - 2\delta_{k,k'} \left[\left\langle \boldsymbol{r}_\ell \,\middle|\, \hat{\boldsymbol{\Phi}} \right\rangle - t_\ell\right] z_k^{(1)} \right\}. \tag{15}$$

Demonstrate that if $\{\lambda_k\}$ and $\{|\lambda_k\rangle\}$ are the eigenvalues and eigenvectors of the Hessian, the solution of the regularized problem is

$$|\tilde{\boldsymbol{w}}_\nu\rangle = \left(\sum_k \frac{\lambda_k}{\lambda_k + \nu} \, |\lambda_k\rangle \langle \lambda_k| \right) |\boldsymbol{w}\rangle \,. \tag{16}$$

The demonstration of equation (16) should be placed in the Methods section of your report, the derivation of the entries of the Hessian, in the Results section.

6. Finally, you are required to present a table with the last updated values of $w_{k,n}^{(1)}$. Observe that the total number of weights should be 60. The only purpose of this task is to verify that you have not copied your results from sombody else (observe that two independent runs of the same program, on the same machine should produce similar estimates for the coefficients $\overline{\boldsymbol{\Phi}}$ but different values for the weights $\boldsymbol{w}$).