Intelligence Artificielle & Apprentissage
Calais ING2/ING3
Questions de révision

Augustin Cosse
augustin.cosse@univ-littoral.fr

January 2022

Question 1 *We consider the neural network shown in Fig. 4 which consists of alternating* 2 *units and* 1 *unit hidden layers. The weights associated to the* i^{th} *unit in layer* k *are denoted as* $w_{ij}^{(k)}$ *and each neuron is equipped with a sigmoid activation and a bias* $w_{i0}^{(k)}$ *(not represented on the Figure)*

1. [1*pts*] *Sketch the sigmoid activation*
2. [2*pts*] *Give the detailed expression of* $y(\boldsymbol{x}; W)$ *as a function of* $\boldsymbol{x}$*, and the* $w_{ij}^{(k)}$*.*
3. [4*pts*] *Using backpropagation, derive the gradient with respect to* $w_{11}^{(1)}$ *for a general* t *and* x *(give all the steps)*

Question 2 *We consider the logistic regression classifier*

$$p(t(\boldsymbol{x}) = 1|\boldsymbol{x}) = \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$
$$p(t(\boldsymbol{x}) = 0|\boldsymbol{x}) = 1 - \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

where $\sigma(x)$ *denotes the usual sigmoid function. Given the data shown in Fig. 2,*

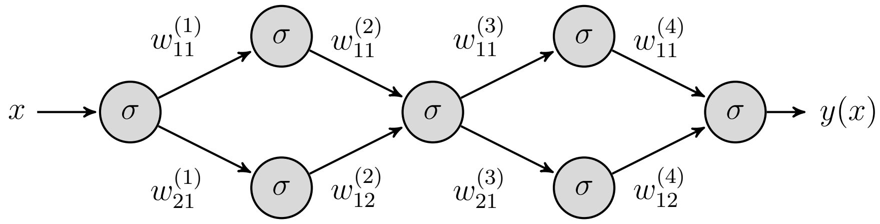

Figure 1: Neural Network for Question 1

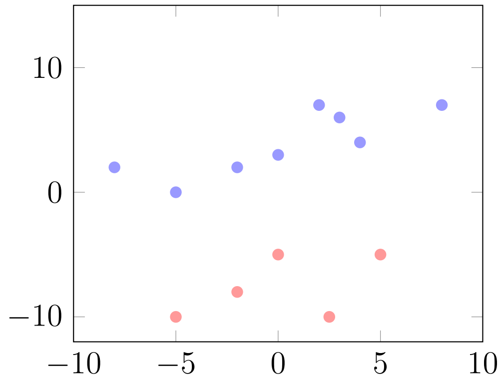

Figure 2: Training set for Question 2.

1. [2pts] *What would be a good choice for the parameters $\beta_0, \beta_1, \beta_2$ (the choice does not need to be optimal)*

2. [2pts] *Let us assume that your solution corresponds to the minimum of a certain loss $\ell(\boldsymbol{\beta})$. How would this solution change if we now decided to minimize $\ell + \lambda R(\boldsymbol{\beta})$ where R denotes the Ridge regularizer. Motivate your answer.*

Question 3 *Give the pseudo-code for the one-vs-rest classifier.*

Question 4 *Consider a real valued feature vector $\boldsymbol{x} = (x_1, x_2, \dots, x_D)$ and real variable t. The t variable is generated, conditional on $\boldsymbol{x}$, from the following process*

$$\varepsilon \sim N(0, \sigma^2)$$
$$t = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon$$

where every ε is an independent variable which is drawn from a Gaussian distribution with mean 0 and standard deviation σ. The conditional distribution of t given $\boldsymbol{x}$ reads as

$$p(t|\boldsymbol{x}, \beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(t - \beta_0 - \boldsymbol{\beta}^T\boldsymbol{x})^2\right)$$

In class we have assumed that the noise variance σ^2 was known. However, we can also use the principle of Maximum Likelihood Estimation to obtain the Maximum Likelihood Estimator (MLE) for the noise variance σ^2_{ML}. To find the expression of σ^2_{ML}, follow the steps below.

1. [2pts] *Start by writing the log-likelihood (taking all the pairs $\left\{\boldsymbol{x}_i, t^{(i)}\right\}_{i=1}^N$ into account)*

2. [2pts] *Compute the derivative of this function with respect to σ^2, set it to 0 and solve the resulting equation*

Question 5 [6*pts*] *Consider real valued variables x and t. The t variable is generated, conditional on x, from the following process*

$$\varepsilon \sim N(0, \sigma^2)$$
$$t = \beta x + \varepsilon$$

where every ε is an independent variable which is drawn from a Gaussian distribution with mean 0 and standard deviation σ. This is a one feature linear regression model, where β is the only weight parameter. The conditional probability distribution of t is given by

$$p(t|x, \beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(t - \beta x)^2\right)$$

1. [2*pts*] *Assume we have a training dataset of n pairs $(x^{(i)}, t^{(i)})$ for $i = 1, \ldots, n$ and σ is known. Which of the following equations correctly represent the maximum likelihood problem for estimating β? (Say yes or no to each possibility, keeping in mind that several of them might be right)*

$$\underset{\beta}{\text{argmax}} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(t^{(i)} - \beta x^{(i)})^2)$$
$$\underset{\beta}{\text{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(t^{(i)} - \beta x^{(i)})^2)$$
$$\underset{\beta}{\text{argmax}} \sum_{i=1}^{n} \exp(-\frac{1}{2\sigma^2}(t^{(i)} - \beta x^{(i)})^2)$$
$$\underset{\beta}{\text{argmax}} \prod_{i=1}^{n} \exp(-\frac{1}{2\sigma^2}(t^{(i)} - \beta x^{(i)})^2)$$
$$\underset{\beta}{\text{argmax}} \frac{1}{2} \sum_{i=1}^{n} (t^{(i)} - \beta x^{(i)})^2$$
$$\underset{\beta}{\text{argmin}} \frac{1}{2} \sum_{i=1}^{n} (t^{(i)} - \beta x^{(i)})^2$$

2. [2*pts*] *Derive the maximum likelihood estimator of the parameter β in terms of the training examples $t^{(i)}$ and $x^{(i)}$. (suggestion: start with the simplest form of the problem you found above and use the fact that the maximum/minimum can be found by setting the derivatives to zero)*

3. [2*pts*] *We now consider a prior on β. Assume that $\beta \sim N(0, \lambda^2)$ so that*

$$p_\lambda(\beta) = \frac{1}{\sqrt{2\pi}\lambda} \exp(-\frac{1}{2\lambda^2}\beta^2)$$

We let β_{MLE} and β_{MAP} denote the Maximum Likelihood and Maximum A Posteriori estimators. Complete the table below

x_1	x_2	$y(x_1, x_2)$
1	1	0
0	0	0
1	0	1
0	1	0

Table 1: Dataset used for Question 6

	$p_\lambda(\beta)$: *wider/narrower/same ?*	$\lvert\beta_{MLE} - \beta_{MAP}\rvert$ *increase/decrease?*
As $\lambda \to \infty$		
As $\lambda \to 0$		

Question 6 (8pts)

1. [5*pts*] *Consider a neural network with two hidden layers:* $d = 2$ *dimensional inputs,* 2 *units in the first hidden layer,* 2 *units in the second hidden layer and a single output.*
   a) *Draw a picture of the network*
   b) *Write out an expression for* $y(x)$ *assuming* $ReLU$ *activation functions. Be as explicit as possible.*
   c) *How many parameters are there?*
2. [3*pts*] *Consider the dataset given in table 1. Can this boolean function be represented by a single neuron with logistic activation function? If yes, give the value of the weights. If not motivate your answer with a short sentence.*

Question 7 *We consider a two hidden layers neural network* $y(\boldsymbol{x}; W)$, $\boldsymbol{x} \in \mathbb{R}^2$ *with a final sigmoid activation (output unit). The first hidden layer consists of* 3 *units and the second hidden layer consists of* 2 *units. The weights from the first and second layers (including the intercepts) are respectively stored in the matrices* $W_1 \in \mathbb{R}^{3\times 3}$ *and* $W_2 \in \mathbb{R}^{2\times 4}$. *The weights associated to the output unit are stored in the vector* $w_{\text{out}} \in \mathbb{R}^3$. *All the hidden units have* $ReLU$ *activations*

1. [2*pts*] *Sketch the* $ReLU$ *and sigmoid functions*
2. [2*pts*] *Sketch the network*
3. [2*pts*] *Give the detailed expression of* $y(\boldsymbol{x}; W)$ *as a function of* $\boldsymbol{x}$, W_1, W_2 *and* w_{out}.

Question 8 *We consider the dataset shown in Fig. 3. Draw on top of this dataset the least squares classifier and the logistic regression classifier. Briefly motivate your answer.*

Question 9 *Describe the backpropagation steps (be as exhaustive as possible)*

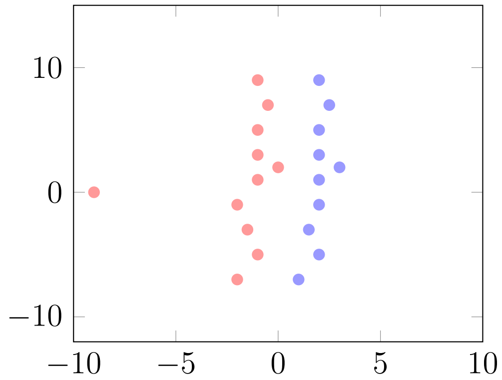

Figure 3: Training set for Question 8.

Question 10 *Consider a neural network with three layers including an input layer. The first (input) layer has four inputs x_1, x_2, x_3 and x_4. The second layer has six hidden units corresponding to all pairwise multiplications. The output node o simply adds the values in the six hidden units. Let L be the loss at the output node. Suppose that you know that $\frac{\partial L}{\partial o} = 2$ and $x_1 = 1$, $x_2 = 2$, $x_3 = 3$ and $x_4 = 4$. Compute $\frac{\partial L}{\partial x_i}$ for each i*

Question 11 *Derive a gradient descent algorithm that minimizes the sum of squared errors for a variant of a perceptron (i.e. one neuron) where the output y of the unit depends on its inputs x_i as follows*

$$y(\boldsymbol{x}) = w_0 + w_1 x_1 + w_1 x_1^3 + w_2 x_2 + w_2 x_2^3 + \ldots + w_n + w_n x_n^3$$

Give your answer in the form $w_i \leftarrow w_i + \ldots$ for $1 \leq i \leq n$.

Question 12 *You want to perform a classification task. You are hesitant between two choices: Approach 1 and Approach 2. The only difference between these two approaches is the loss function that is minimized. Assume that $x^{(i)} \in \mathbb{R}$ and $t^{(i)} \in \{+1, -1\}$, $i = 1, \ldots, m$ are the i^{th} example and output label in the dataset, respectively. $f(x^{(i)})$ denotes the output of the classifier for the i^{th} example. Recall that for a given loss ℓ, you minimize the cost*

$$J = \frac{1}{m} \sum_{i=1}^{n} \ell(f(x^{(i)}), t^{(i)}) \tag{1}$$

As we mentioned, the only difference between approach 1 and approach 2 is the choice of the loss function:

$$\ell_1(f(x^{(i)}), t^{(i)}) = \max\left\{0, 1 - t^{(i)} f(x^{(i)})\right\} \tag{2}$$
$$\ell_2(f(x^{(i)}), t^{(i)}) = \log_2(1 + \exp(-t^{(i)} f(x^{(i)}))) \tag{3}$$

1. *Rewrite ℓ_2 in terms of the sigmoid function.*
2. *You are given an example with $t^{(i)} = -1$. What value of $f(x^{(i)})$ will minimize ℓ_2?*

3. *Assume that an outlier (very far from the decision boundary but in the right class) is added to the dataset. How will that affect classifier* (2)*? Why?*

4. *You are given an example with* $t^{(i)} = -1$. *What is the greatest value of* $f(x^{(i)})$ *that will minimize* ℓ_1*?*

5. *You would like a classifier whose output can be interpreted as a probability. Which loss function is better and why?*

Question 13 *Indicate whether the following statements are true or false*

Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(t^{(i)} - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s$$

True / False	*As we increase* s *from* 0*, the training RSS will increase initially, and then eventually start decreasing in an inverted U-shape*
True / False	*As we increase* s *from* 0*, the training RSS will decrease initially, and then eventually start increasing in an inverted U-shape*
True / False	*As we increase* s *from* 0*, the training RSS will steadily increase*
True / False	*As we increase* s *from* 0*, the training RSS will steadily decrease*

Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(t^{(i)} - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for a particular value of λ

True / False	*As we increase* λ *from* 0*, the variance will increase initially, and then eventually start decreasing in an inverted U-shape*
True / False	*As we increase* λ *from* 0*, the variance will decrease initially, and then eventually start increasing in an inverted U-shape*
True / False	*As we increase* λ *from* 0*, the variance will steadily increase*
True / False	*As we increase* λ *from* 0*, the variance will steadily decrease*

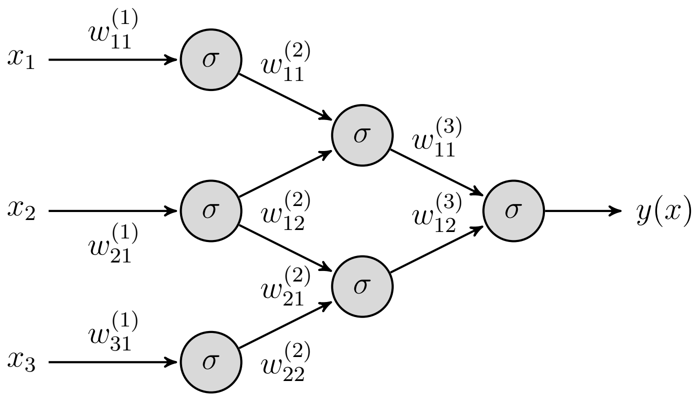

Figure 4: Neural Network used for question 14

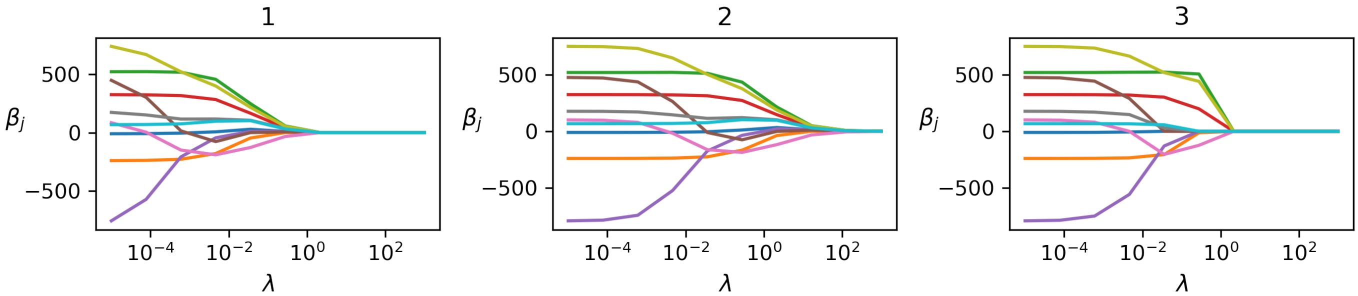

Figure 5: Evolution of the regression coefficients for an increasing value of the regularization weights λ_1, λ_2 in the case of the elastic net model. The various lines correspond to different regression coefficients β_j.

Question 14 *We want to use the backpropagation algorithm, in order to compute the gradient of the binary cross entropy loss (for a single pair $(\boldsymbol{x}^{(i)}, t^{(i)})$) with respect to the weight $w_{11}^{(1)}$ for the network shown in Fig. 4. To do so, we will proceed as follows:*

1. [1pts] *Give the expression of the binary cross entropy loss for the pair $\left\{\boldsymbol{x}^{(i)}, t^{(i)}\right\}$*

2. [1pts] *Give the expression of $\delta^{(3)} = \delta_{out} = \frac{\partial L}{\partial a_{out}}$ (derivative of the binary cross entropy loss with respect to the output pre-activation)*

3. [2pts] *Give the backpropagation equation and use this equation to derive, from δ_{out}, the values of the δ_i^2 for $i = 1, 2$. Then, from the δ_i^2, obtain the value of δ_1^1.*

4. [1pts] *Finally, give the expression of the derivative $\frac{\partial L}{\partial w_{11}^1}$ as a function of δ_1^1 and $z_1^{(0)} = x_1$. Deduce from this, and from your expression for δ_1^1, the final answer to the question.*

Question 15 *We consider the following regression model, known as "elastic net regularization"*

$$L\left(\beta, \{\boldsymbol{x}^{(i)}, t^{(i)}\}_{i=1}^{N}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - \beta_0 - \sum_{j=1}^{D}\beta_j x_j^{(i)}\right)^2 + \lambda_2\left(\sum_{j=1}^{D}|\beta_j|^2\right) + \lambda_1\left(\sum_{j=1}^{D}|\beta_j|\right) \quad (4)$$

1. [1*pt*] *Indicate the differentiable and non-differentiable parts of the loss.*

2. [2*pts*] *Figure 5 illustrates the evolution of the regression coefficients (each of the* β_j *is represented by a different curve) obtained by minimizing the loss* (4) *for different choices of* (λ_1, λ_2)*. In particular, the figure illustrates each of the following scenarios:*

   - *Ridge regularization (*$\lambda_2 > 0, \lambda_1 = 0$*)*
   - *LASSO regularization (*$\lambda_1 > 0, \lambda_2 = 0$*)*
   - *A trade-off between Ridge and LASSO corresponding to non zeros* λ_1 *and* λ_2*, with* $\lambda_1 = 9\lambda_2$

   Indicate, on each of the subfigures, the model to which it corresponds.