CSCI-UA 9473 - Introduction to Machine Learning
Midterm

Augustin Cosse

October 2022

Total: 35 points
Duration: 3h
General instructions: The exam consists of 2 questions (each question consisting itself of several subquestions). Once you are done, make sure to write your name on each page, then take a picture of all your answers and send it by email to acosse@nyu.edu. In case you have any question, you can ask those through the chat. Answer as many questions as you can starting with those you feel more confident with.

Question 1 (15pts)

1. [5pts] *Indicate whether the following statements are true or false*

True / False	*The derivative of the sigmoid function satisfies* $\sigma'(x) = \sigma(x)(1-\sigma(x))$
True / False	*The derivative of the sigmoid function satisfies* $\sigma'(x) = \sigma(x)(\sigma(x)-1)$
True / False	*Once learned, the linear discriminant analysis model can be reduced to a logistic regression classifier*
True / False	*Linear Discriminant Analysis with a diagonal covariance can be considered an instance of a Naive Bayes classifier*
True / False	*In cross validation, the whole dataset is used for testing and for training*
True / False	*More complex models will have a larger variance contribution to the MSE than simpler models*
True / False	*Simpler models will have a larger bias contribution to the MSE than more complex models*
True / False	*If we consider the linear regression problem with the regularizer given in* (1), *larger values of* p *are more likely to lead to an increase in the number of vanishing coefficients than smaller values*

$$\mathcal{R}(\beta) = \left(\sum_{j=1}^{D} |\beta_j|^p\right)^{1/p} \tag{1}$$

2. [3pts] *What is the difference between generative and discriminative classifiers? Give one example of a classifier from each family.*

3. [4pts] *Consider a real valued feature vector* $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)$ *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.

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

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

4. [3pts] *Give the pseudo-code for the one-vs-rest classifier.*

Question 2 (20pts)

1. [5pts] *Indicate whether the following statements are true or false*

True / False	*Gradient descent on the Ridge loss will always converge to the global minimum of the loss*
True / False	*Increasing the number of units in the hidden layer of a one hidden layer neural network will increase the variance*
True / False	*A multi-layer network that uses only the identity activation function in all its layers reduces to a single-layer network performing linear regression*
True / False	*When training a deep neural network, using a sufficiently small learning rate and a sufficiently large number of iterations guarantees that the optimizer will find the global minimum*
True / False	*Neural networks can be considered as parametric models*
True / False	*When designing neural networks, it is always better to use the sigmoid activation to avoid the vanishing gradient problem*

2. [4pts] *We consider a simple regression model with two coefficients* $t = \beta_1 \overline{x}_1^s + \beta_2 \overline{x}_2^s$. *We assume that the data has been centered so that the model is learned on* $\overline{x}^{(i)} = x^{(i)} - \frac{1}{N}\sum_i x^{(i)}$ *and* $\overline{t}^{(i)} = t^{(i)} - \frac{1}{N}\sum_i t^{(i)}$. *moreover, after the centering step, the* $\overline{x}^{(i)}$ *are scaled as*

$$\overline{x}_k^{s,(i)} = \overline{x}_k^{(i)} \leftarrow \overline{x}_k^{(i)}/\sigma_k$$

where σ_k^2 *is the variance associated to the* k^{th} *feature of* $\boldsymbol{x}^{(i)}$,

$$\sigma_k^2 = \frac{1}{N}\sum_{i=1}^{N}(x_k^{(i)} - \overline{x}_k)^2, \quad \overline{x}_k = \frac{1}{N}\sum_i x_k^{(i)}$$

   (a) [2pts] *Show that the normal equations in this case can read as*

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}$$

   What is the expression for r_{12} *in terms of the original* $x_1^{(i)}, x_2^{(i)}$*? (Start by writing the expression of* r_{12} *as a function of the* $\overline{x}^{s,(i)}$ *then replace the* $\overline{x}^{s(i)}$ *by their expression as a function of the* $x^{(i)}$*)*

   (b) [2pts] *Give the expression of the inverse* $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ *as a function of* r_{12}. *What are the values of* r_{12} *for which this inverse is well defined?*

3. [7pts] *We consider the neural network shown in Fig. 2 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)*

   (a) [1pts] *Sketch the sigmoid activation*

   (b) [2pts] *Give the detailed expression of* $y(\boldsymbol{x}; W)$ *as a function of* $\boldsymbol{x}$, *and the* $w_{ij}^{(k)}$.

   (c) [4pts] *Using backpropagation, derive the gradient with respect to* $w_{11}^{(1)}$ *for a general* t *and* x *(give all the steps)*

4. [4pts] *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. 1,*

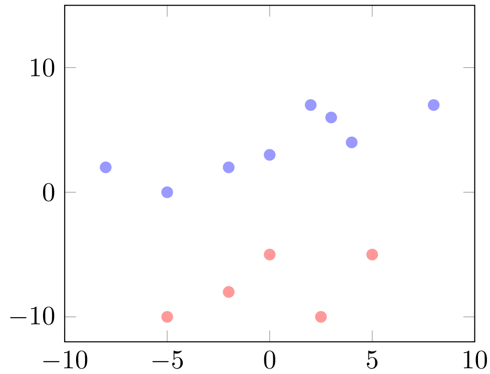

Figure 1: Training set for Question 2.4.

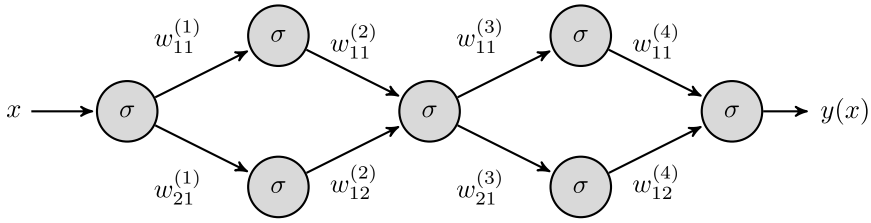

Figure 2: Neural Network for Question 2.3

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

(b) [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.*