CSCI-UA 9473 - Introduction to Machine Learning
Midterm

Augustin Cosse

April 2022

Total: 45 points
Total time: 1h15
General instructions: The exam consists of 3 questions (each question consisting itself of 3 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 (Regression and regularization 15pts)

1. *Indicate whether the following statements are true or false (5pts)*

True / False	*Gradient descent finds the global minimum of the least squares loss for linear regression*
True / False	*Gradient descent finds the global minimum of the ridge loss for linear regression*
True / False	*Least squares regression can be understood as a Maximum a Posteriori (MAP) approach with a uniform prior on the regression weights*
True / False	*Least squares regression can be understood as a Maximum Likelihood (MLE) approach with Gaussian prior on the deviations* $t^{(i)} - \left(\beta_0 + \sum_{p=1}^{D} \beta_p x_p\right)$
True / False	*LASSO regression can be understood as a Maximum a Posteriori (MAP) approach with a Poisson prior on the regression weights*
True / False	*Ridge regression can be understood as a Maximum a Posteriori (MAP) approach with a Gaussian prior on the regression weights*
True / False	*The Ridge minimizer corresponds to increasing the eigenvalues of* $\boldsymbol{X}^T\boldsymbol{X}$ *by* λ

2. *We consider the simple dataset show in Fig 1 below. We want to learn a linear regression model on the points shown in red. Explain how you would proceed (all steps + pseudo-code) [7pts]*

3. *The balls* ($\|\boldsymbol{\beta}\|_p^p = \left(\sum_{k=1}^{D} |\beta_k|^p\right) \leq 1$) *corresponding to the LASSO* ($p = 1$) *and ridge* ($p = 2$) *formulations are shown in Fig. 2 below. Sketch the* ℓ_p *balls for* $p < 1$ *and* $p > 2$ *[3pts]*

Question 2 (Neural network 15pts)

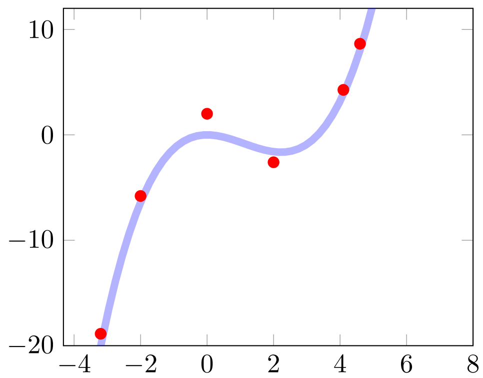

Figure 1: Training set for Question 1. The blue curve corresponds to the equation $0.3x^3 - x^2$ and the red points are respectively located at $(0, 2)$, $(2, -2.6)$, $(-2, -5.8)$, $(-3.2, -18.8704)$ $(4.1, 4.2663)$, $(4.6, 8.6408)$

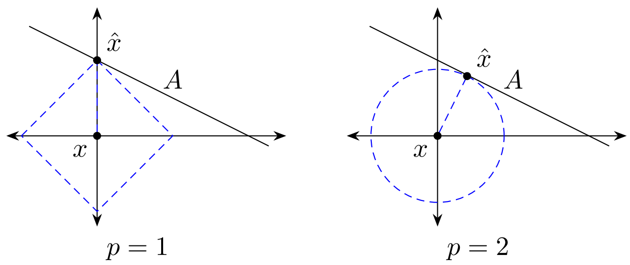

Figure 2: l_p balls for $p = 1$ (LASSO) and $p = 2$ (ridge)

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

True / False	*Neural Networks cannot model the XOR dataset*
True / False	*Neural networks cannot be used with the binary cross entropy loss*
True / False	*Neural networks cannot be used in regression*
True / False	*Increasing the number of hidden layers in a network will increase the bias*
True / False	*Increasing the number of hidden layers in a network will increase the variance*

2. *Describe the backpropagation steps (be as exhaustive as possible, no need to provide a python implementation) [5pts]*

3. *We consider the dataset shown in Fig. 3. Explain how you would build a neural network for this dataset (including number of hidden layers and activation functions) and draw the separating planes on top of the dataset. [5pts]*

Question 3 (Kernels, 15pts)

1. *Explain how we can learn a classifier based on the Gaussian kernel with the least squares loss (all the steps + pseudo code) [5pts]*

2. *When is a kernel considered valid (give a formal criterion)? We consider the kernel $\kappa(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ defined as $\kappa(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) = \left((\boldsymbol{x}^{(i)})^T(\boldsymbol{x}^{(j)}) + c\right)^2$ where c is a positive constant. Is this a valid kernel? Motivate your answer with a proof. [5pts]*

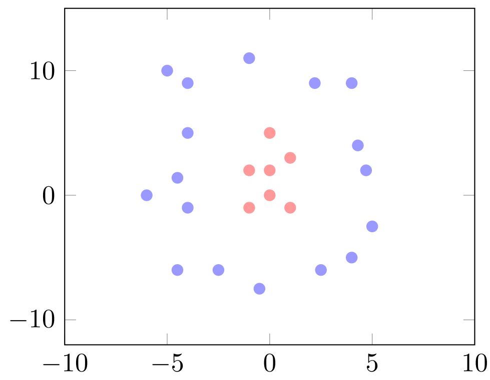

Figure 3: Training set for Question 2.

3. *We consider the dataset shown in Fig 4. we assume that the red points have target* $+1$ *and the blue ones have target* -1. *On that dataset, we want to learn a kernel classifier of the form*

$$y(\boldsymbol{x}) = \sum_{i=1}^{N} \lambda_i \kappa(\boldsymbol{x}, \boldsymbol{x}^{(i)})$$

with a Gaussian kernel, $\kappa(\boldsymbol{x}, \boldsymbol{x}^{(i)}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}^{(i)}\|^2}{\sigma}\right)$. *What could be appropriate values for* λ_i, σ *(Motivate your answer, associate each lambda* λ_i *to each point* $\boldsymbol{x}^{(i)}$ *from the dataset by superimposing the* λ_i *on their respective data points in Fig 4). Plot the resulting classifier on top of the data. [5pts]*

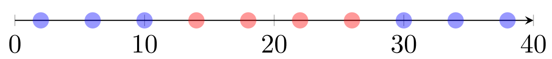

Figure 4: Training set for Question 3.