CSCI-UA 9473 - Introduction to Machine Learning
Midterm retake

Augustin Cosse

July 2022

Total: 34 points
Total time: 2h00
General instructions: The exam consists of 2 questions (each question consisting itself of multiple 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 (Part I 18pts)

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

True / False	*Gradient descent can be applied to the function* $f(x_1, x_2) = x_1^2 + x_2^2 + \|3x_1x_2 - 2\|$
True / False	*Gradient descent will always decrease the value of a loss provided that the learning rate is sufficiently small and the gradient is non zero*
True / False	*When trying to learn a regression model on feature vectors for which we know that only some of the features are useful, we should use a LASSO formulation*
True / False	*In Linear Discriminant Analysis, the Gaussian distribution is used to model the class probability* $p(\boldsymbol{x}\|t)$
True / False	*The total number of parameters of a Linear Discriminant Analysis model used for a binary classification problem is 5*
True / False	*In linear regression, given an infinite number of (noiseless) training examples, we can expect the training error to go to zero*
True / False	*When using gradient descent, if* $\boldsymbol{\beta}$ *is initialized at a local minimum, the regression coefficients will not get updated*

2. [5pts] *We consider the logistic regression classifier* $y(\boldsymbol{x}) = \sigma(\boldsymbol{\beta}^T\tilde{\boldsymbol{x}})$ *where* $\tilde{\boldsymbol{x}} = [1, \boldsymbol{x}^T] \in \mathbb{R}^{D+1}$, $\boldsymbol{\beta} = [\beta_0, \ldots, \beta_D] \in \mathbb{R}^{D+1}$ *and* σ *is the sigmoid activation.*
   (a) [1pt] *Is logistic regression a generative or a discriminative classifier?*
   (b) [1pt] *Give the expression as well as a sketch of the sigmoid activation* $\sigma(x)$
   (c) [1pt] *Give the expression of the log loss* ℓ *(corresponding to the minimization of the negative log-likelihood) for the logistic regression classifier.*
   (d) [2pts] *Show that the log-loss satisfies*

$$\begin{aligned}\text{grad } \ell(\beta) &= -\frac{1}{N}\sum_{i=1}^{N}\frac{t^{(i)}\tilde{\boldsymbol{x}}^{(i)}}{1+e^{t^{(i)}\boldsymbol{\beta}^T\tilde{\boldsymbol{x}}^{(i)}}}\\ &= \frac{1}{N}\sum_{i=1}^{N} -t^{(i)}\tilde{\boldsymbol{x}}^{(i)}\sigma(-t^{(i)}\boldsymbol{\beta}^T\tilde{\boldsymbol{x}}^{(i)})\end{aligned}$$

Table 1: Dataset for Question 1.3

x_1	x_2	x_3	t
0	0	0	1
0	1	0	0
0	1	1	1
1	1	1	0

3. [*4pts*] *We want to learn a binary (linear) classifier on the dataset given in Table 1 above.*
   (a) [*2pts*] *Give the set of linear inequalities the weights $\beta_1, \beta_2, \beta_3$ and bias β_0 have to satisfy.*
   (b) [2*pts*] *Give a setting of the weights and bias that correctly classifies the training examples from Table 1.*

4. [*4pts*] *We consider the regression problem shown in Fig 1. Assuming that the complexity of the model is unknown, we consider a degree-10 model including all the features $1, x, x^2, \ldots, x^{10}$.*
   (a) [*3pts*] *Explain how to use Best Subset Selection to recover the true model.*
   (b) [*1pts*] *Sketch the evolution of the cross validation error as a function of the number of features for this particular problem.*

Question 2 (Part II 16pts)

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

True / False	*A family of models learned by minimizing the Ridge loss with $\lambda > 0$ will have smaller variance than a model learned by minimizing the residual sum-of-squares loss*
True / False	*The XOR dataset can be learned by a neural network with a single hidden layer*
True / False	*The perceptron algorithm will converge in a finite number of steps on any dataset that is linearly separable (provided the learning rate is small enough)*
True / False	*The solution resulting from the minimization of the Ridge loss is $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{t}$*
True / False	*Minimizing the LASSO corresponds to looking for the Maximum-A-Posteriori estimator with a Laplace prior*
True / False	*A neural network can be trained by minimizing the log loss or the residual sum-of-squares loss*

2. [*5pts*] *Describe the backpropagation steps (be as exhaustive as possible)*

3. [*3pts*] *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*

4. [*3pts*] *Explain why the kernel trick allows us to solve a learning problem (e.g. a regression problem) in a high dimensional feature space without significantly increasing the running time.*

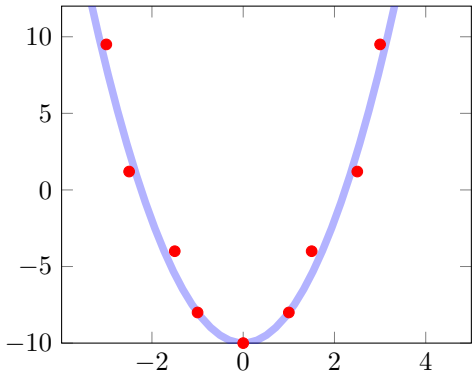

Figure 1: Training set for Question 1.4. The blue curve corresponds to the equation $y(x) = 2x^2 - 10$.