Final Exam CSCI-UA9473 - Intro to Machine Learning
Fall 2022

December 12, 2022

Name :

Total: 31 points
Duration: 3h
General Instructions: The exam consists of two main parts (Each of those parts containing multiple subquestions). You are free to write your answers on supplementary pages but you should make sure to clearly indicate your name, as well as the number of the question on each additional page. Answer as many questions as you can, starting with those you are more comfortable with.

Question 1 (14pts)

1. [5pts] *For each of the following statements, indicate whether the statement is true or false.*

True / False	*In the setting of binary classification, minimizing the binary cross entropy is equivalent to computing a maximum likelihood estimator*
True / False	*In classical PCA, the matrix encoding the latent subspace is not uniquely defined*
True / False	*When considering Ridge regression, and for a regularization weight $\lambda > 0$, increasing the value of λ will result in an increase of the variance of the corresponding family of models*
True / False	*When considering linear regression, adding a Ridge penalty, with associated weight $\lambda > 0$, will result in a translation of the eigenvalues of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ by λ*
True / False	*The normal equations always have at least one solution*
True / False	*When learning a linear classifier through gradient descent, initializing the weights and biases to zero will prevent any update in those parameters*
True / False	*A maximum margin separating hyperplane can be learned by minimizing the norm of the normal vector to the hyperplane under a set of linear constraints*

2. [3pts] *Give the pseudo-code for "K-means" (including the initialization)*

3. [6pts] *We consider the following regression model, known as "elastic net regularization"*

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

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

(b) [2pts] *Figure 2 illustrates the evolution of the regression coefficients (each of the* β_j *is represented by a different curve) obtained by minimizing the loss* (1) *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.

(c) [3pts] *We consider the projector* $\mathcal{T}_{\lambda\eta}(\boldsymbol{\beta})$ *whose* i^{th} *component when applied to a weight vector* $\boldsymbol{\beta}$ *is defined as*

$$[\mathcal{T}_{\lambda\eta}(\boldsymbol{\beta})]_i = \max(0, |\beta_i| - \lambda\eta)\text{sign}(\beta_i) \tag{2}$$

The projector $\mathcal{T}_{\lambda\eta}$ *therefore replaces by* 0 *all the regression coefficients* β_i *whose magnitude is smaller than* $\lambda\eta$. *Relying on this projector, provide a minimization algorithm for the loss* (1).

Question 2 (17pts)

1. [5pts] *For each of the following statements, indicate whether the statement is true or false.*

True / False	*Increasing the number of neurons in the hidden layer of a one hidden layer neural network will increase the variance of the corresponding family of models*
True / False	*Single linkage clustering merges, at each step, the two clusters that minimize the distance between their closest two points*
True / False	*In the A priori algorithm, the support of a rule '$A \Rightarrow B$' can be interpreted as the total number of transactions including all the items in A and the items in B*
True / False	*The smallest number of neurons needed to learn the XOR model is equal to 4*
True / False	*In terms of expressivity (i.e. the ability of a network to capture a particular data distribution), a neural network is more powerful than a linear model based on polynomial features*

2. [3pts] *We consider the neural network shown in Fig. 1, which includes 3 hidden layers. The weights associated to unit i from layer ℓ are encoded by the variables $w_{ij}^{(\ell)}$. Each neuron is defined with an associated sigmoid activation σ, and a bias $w_{i0}^{(\ell)}$ (not represented on the figure)*

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

   (b) [2pts] *Give the expression of $y(\boldsymbol{x})$ as a function of $\boldsymbol{x} = (x_1, x_2, x_3)$, the weights $w_{ij}^{(\ell)}$ and the biases $w_{i0}^{(\ell)}$.*

3. [5pts] *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. 1. To do so, we will proceed as follows:*

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

   (b) [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)*

   (c) [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.*

   (d) [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.*

4. [4pts] *Let $\boldsymbol{x} = (x_1, x_2, \ldots x_D), \boldsymbol{y} = (y_1, y_2, \ldots, y_D)$. We consider the kernel $\kappa(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y} + c)^2$ where $c > 0$ is a positive constant.*

   (a) [2pts] *Is this kernel a valid kernel ? Motivate your answer with a short proof.*

   (b) [2pts] *How about $\tilde{\kappa}(\boldsymbol{x}, \boldsymbol{y}) = \kappa(\boldsymbol{x}, \boldsymbol{y}) + \cos(x_1 - y_1)$?*

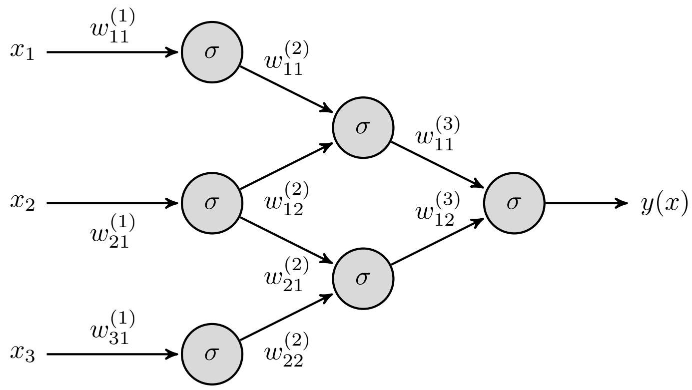

Figure 1: Neural Network used for questions 2.2 and 2.3

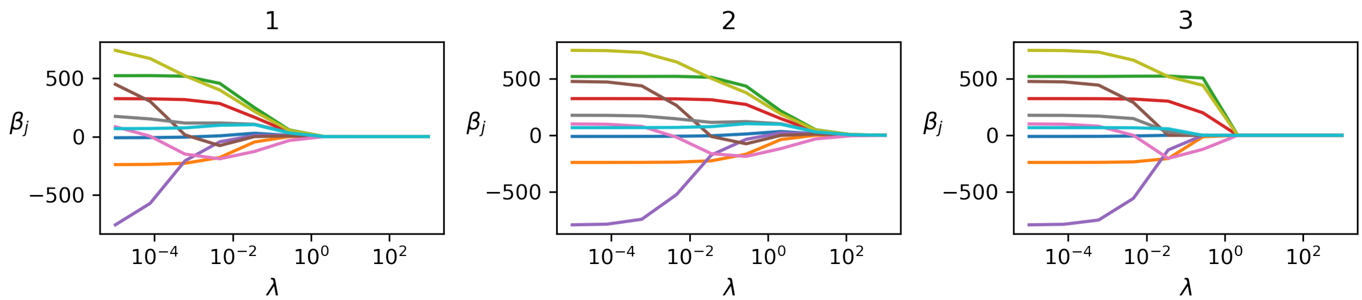

Figure 2: Evolution of the regression coefficients for an increasing value of the regularization weights λ_1, λ_2 in the case of the elastic net model. The various lines correspond to different regression coefficients β_j.