CSCI-UA 9473 - Introduction to Machine Learning
Final III

Augustin Cosse

May 2022

Total: 45 points
Total time: 2h00
General instructions: The exam consists of 2 parts, a first part focusing on supervised learning (including 5 questions), and a second part focusing on unsupervised learning (including 5 questions). 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 (Supervised Learning 25pts)

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

True / False	*LASSO is a parametric method*
True / False	*Suppose that we have a regularized linear regression model* $\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\|t-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$. *Increasing* λ *will increase the variance and decrease the bias*
True / False	*Suppose that we have a regularized linear regression model* $\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\|t-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1$. *Increasing* λ *will increase the bias and decrease the variance*
True / False	*Ridge regression minimizes the squared* ℓ_2 *norm as a penalty on the regression weights*
True / False	*Ridge regression minimizes the* ℓ_2 *norm as a penalty on the regression weights*
True / False	*At the* i^{th} *iteration of online perceptron learning, you have a model* $h_{\boldsymbol{\beta}}$ *and you receive a new instance* $\boldsymbol{x}^{(i+1)}$. *You find out that your current model misclassifies the instance as* $h_{\boldsymbol{\beta}^{(i)}}(\boldsymbol{x}^{(i+1)})=+1$ *when the actual label is* $t^{(i+1)}=-1$. *You update the model using the perceptron algorithm and get a new classifier* $h_{\beta^{(i+1)}}$. $h_{\beta^{(i+1)}}$ *is guaranteed to classify* $\boldsymbol{x}^{(i+1)}$ *correctly as* -1
True / False	*You are using a Maximum Margin classifier with a Gaussian kernel defined as* $e^{-\frac{\|\boldsymbol{x}\|_2^2}{\sigma^2}}$ *for a classification problem. You find that the training accuracy is* 0.97 *but the test accuracy is* .65. *The test accuracy can be improved by increasing the kernel width* σ
True / False	*You are using a Maximum Margin classifier with a Gaussian kernel defined as* $e^{-\frac{\|\boldsymbol{x}\|_2^2}{\sigma^2}}$ *for a classification problem. You find that the training accuracy is* 0.97 *but the test accuracy is* .65. *The test accuracy can be improved by decreasing the kernel width* σ
True / False	*Consider the dataset shown in Fig. 1. This dataset can be made separable by a linear SVM using only two support vectors.*

True / False *The LASSO formulation is more likely to to set some of the regression weights to zero than is Ridge*

2. [6*pts*] *You have a single hidden-layer neural network for a binary classification task. The input is* $\boldsymbol{x} \in \mathbb{R}^D$, *output* $y \in \mathbb{R}$ *and the true label* $t \in \mathbb{R}$. *The forward propagation equations are*

$$a^{[1]} = W^{[1]}\boldsymbol{x} + b \tag{1}$$

$$z^{[1]} = \sigma(a^{[1]}) \tag{2}$$

$$y(\boldsymbol{x}^{(i)}) = z^{[1]} \tag{3}$$

$$\ell = -\sum_{i=1}^{m} t^{(i)} \log(y(\boldsymbol{x}^{(i)})) + (1 - t^{(i)}) \log(1 - y(\boldsymbol{x}^{(i)})) \tag{4}$$

   a) *Explain why minimizing the loss* (4) *corresponds to looking for the maximum likelihood estimator of the network.*

   b) *Write the expression for* $\frac{\partial \ell}{\partial W^{[1]}}$ *as a matrix product of two terms.*

   c) *Explain how to extend your result to more hidden layers.*

3. [2*pts*] *Why is it important to place non-linearities between the layers of neural networks?*

4. [5*pts*] *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*

$$J = \frac{1}{m} \sum_{i=1}^{n} \ell(f(x^{(i)}), t^{(i)}) \tag{5}$$

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

$$\ell_1(f(x^{(i)}), t^{(i)}) = \max\left\{0, 1 - t^{(i)} f(x^{(i)})\right\} \tag{6}$$

$$\ell_2(f(x^{(i)}), t^{(i)}) = \log_2(1 + \exp(-t^{(i)} f(x^{(i)}))) \tag{7}$$

   (a) *Rewrite* ℓ_2 *in terms of the sigmoid function.*

   (b) *You are given an example with* $t^{(i)} = -1$. *What value of* $f(x^{(i)})$ *will minimize* ℓ_2*?*

   (c) *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* (6)*? Why?*

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

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

5. [5*pts*] *Give the pseudo code for the one vs rest classifier.*

6. [2*pts*] *We consider the following kernel*

$$k(x, y) = xy + x^2y^2 + \sqrt{xy} + \cos(x - y)$$

   is this a valid kernel? Why?

Question 2 (Unsupervised 20pts)

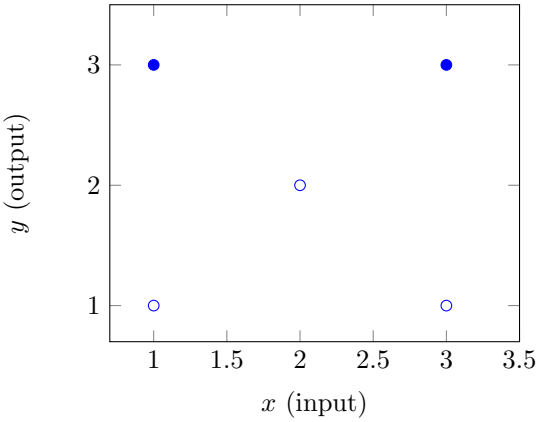

Figure 1: Classification dataset used in Question 1

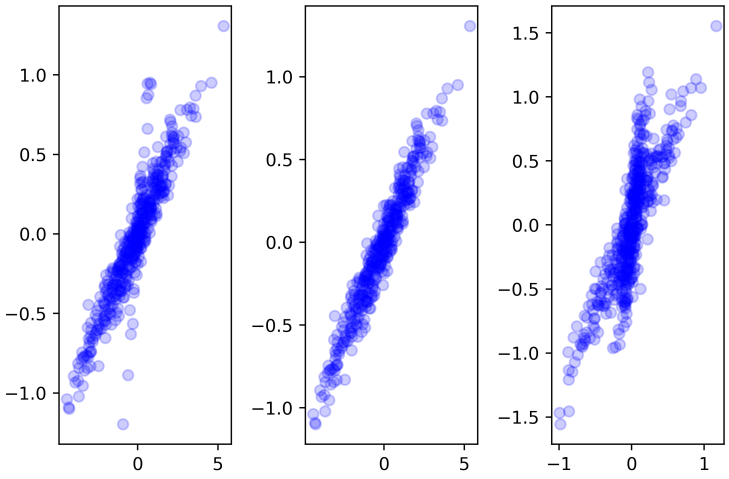

Figure 2: Unsupervised dataset used in Question 2.

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

True / False	*Both PCA and linear regression can be thought of as algorithms for minimizing a sum of squared errors*
True / False	*Implementing K-medoid is more expensive (in terms of computation) than implementing K-means*
True / False	*The A priori algorithm works by discarding item sets whose support is smaller than a given threshold*
True / False	*When merging subclusters, single linkage clustering favors subclusters whose combination will have the smallest diameter*
True / False	*In Factor Analysis, if we write the decomposition into latent factors as* $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z}$*, the factor loading matrix* $\boldsymbol{W}$ *is defined up to a rotation.*
True / False	*Applying Expectation Maximization on a Gaussian Mixture model can be interpreted as a probabilistic version of K-means*

2. *Principal component analysis is a dimensionality reduction method that projects a dataset onto onto its most variable components. You are give the datasets shown in Fig. 2. Draw the first and second components on each plot (if you work with pen and paper, just label the subplots as a, b and c and sketch the corresponding directions on your piece of paper). Explain how one can compute those components. [5pts]*
3. *Give the pseudo-code for K-means [5pts]. What happens when there is an empty cluster?*
4. *Explain how to merge the clusters in agglomerative clustering [3pts].*
5. *Explain the difference between Principal Component Analysis and Factor Analysis [2pts]*