CSCI-UA 9473 - Introduction to Machine Learning
Final II

Augustin Cosse

May 2022

Total: 40 points
Total time: 1h30
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 20pts)

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

True / False	*Gaussian discriminant Analysis can be considered as a discriminative classifier*
True / False	*The MLE estimator can be understood as an MAP estimator with a uniform prior*
True / False	*Minimizing the log loss will always find the MLE estimator in binary classification*
True / False	*The number of parameters in a parametric model is fixed, while the number of parameters in a non-parametric model grows with the amount of training data.*
True / False	*As model complexity increases, bias will decrease while variance will increase*
True / False	*In an algorithm that uses the kernel trick, the Gaussian kernel gives a regression function or prediction function that is a linear combination of Gaussians centered at the sample points*
True / False	*The solution of the regression problem can always be computed as* $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$
True / False	*The solution of the ridge regression problem for* $\lambda > 0$ *can always be computed as* $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

2. *We consider a dataset* $\left\{x^{(i)}, t^{(i)}\right\}_{i=1}^{N}$ *of size* N*. We would like to learn a regression model for this dataset of the form* $y(x^{(i)}) = \beta_0 + \beta_1 x^{(i)} + \beta_2 (x^{(i)})^2$*. We also know that the noise has distribution* $p_\lambda(\varepsilon)$*, that is to say* $t^{(i)} - \beta_0 + \beta_1 x^{(i)} + \beta_2 (x^{(i)})^2 \sim p_\lambda$ *(*p_λ *is a given function parametrized by a (known) parameter* λ*). If we want to use a prior* $h(\beta)$ *on* β_0*,* β_1 *and* β_2*, derive the function that we need to minimize to recover the maximum likelihood estimator (all the samples* $\left\{x^{(i)}, t^{(i)}\right\}$ *and regression coefficients* β_j *are assumed to be independent). Then give the gradient steps. [5pts]*

3. *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 run time. [3pts]*

4. [3*pts*] *Consider the dataset shown in Fig. 1. Circle all the classifiers that will achieve zero training error on this dataset*

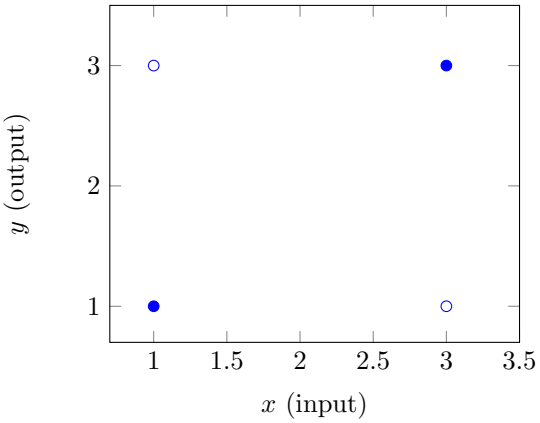

Figure 1: Regression dataset used in Question 1

a) *Logistic regression*

b) *SVM (quadratic kernel, $k(x, y) = (x^T y + c)^2$)*

c) *SVM (Gaussian kernel)*

d) *Perceptron*

e) *Neural network with one hidden layer and two units in the hidden layer (not including the output unit)*

f) *Neural network with one hidden layer and three units in the hidden layer (not including the output unit)*

5. *Derive the expression of the shortest distance from a point $\boldsymbol{z}$ to a hyperplane $\boldsymbol{w}^T\boldsymbol{x}$ (give all the steps and illustrate with a drawing) then deduce from it, the formulation of the Max-Margin classifier [4pts]*

Question 2 (Unsupervised 20pts)

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

True / False	*In PCA, the latent factors are recovered by projecting the prototypes onto the eigenvectors of the sample covariance matrix*
True / False	*The only way to fix the unidentifiability of the factor loading matrix $\boldsymbol{W}$ in FA is to require that matrix to be orthogonal*
True / False	*In the A Priori Algorithm, The confidence, or "predictability" of a rule, is its support divided by the support of its antecedent*
True / False	*When merging subclusters, complete linkage clustering favors subclusters whose combination will have the smallest diameter*
True / False	*The only difference between PCA and FA is that the latent factors in PCA are assumed to follow a Laplace distribution*
True / False	*To work, Expectation Maximization (EM) requires the distributions to be Gaussian*

2. *We consider a data matrix $\boldsymbol{X}$ and we want to learn the best dimension 2 subspace to represent the data. Explain how you would proceed (all details, including pseudo-code)[5pts]*

3. *Give the pseudo-code for the EM algorithm and explain each of the parameters involved [5pts] ?*

4. *Explain how to split the clusters in divisive clustering [3pts].*

5. *Explain the difference between K-means and K-medoid [2pts]*